Multimodal analysis of verbal and nonverbal behaviour on the example of clinical depression

Thesis · May 2015

CITATIONS
16

READS
1,290

1 author:

Sharifa Alghowinem
Massachusetts Institute of Technology
60 PUBLICATIONS **1,497** CITATIONS

SEE PROFILE

Multimodal Analysis of Verbal and Nonverbal Behaviour on the Example of Clinical Depression

Sharifa Alghowinem

A thesis submitted for the degree of
Doctor of Philosophy at
The Australian National University

May 2015

Except where otherwise indicated, this thesis is my own original work.

Sharifa Alghowinem
18 May 2015

to my husband for his patience and support

Acknowledgments

First of all, I thank God for all the peace, comfort and acceptance that my soul needed to reach this point and beyond.

I would like to thank Professor Michael Wagner for opening the door of opportunity to the journey of a PhD, where my main supervisor, Associate Professor Roland Goecke, took me by hand and walked me through the path step by step to the finish line. Roland, thank you for your insights, suggestions and guidance throughout this journey, which made it enjoyable and achievable. I would also like to emphasise thanking you for your understanding, flexibility and patience toward my personal life for all my PhD years. Michael, beside thanking you for your active supervision in my research, I thank you for your support and your friendship. Moreover, I would like to thank the rest of the members of my supervisory panel: Associate Professor Julien Epps, Professor Tom Gedeon, and Dr. Lexing Xie. A special appreciation to Julien for his constructive feedback that improved my prospective toward the research.

My appreciation goes to the Black Dog Institute in Sydney for making my research feasible with their amazing efforts to make this unique multi-disciplinary field available. Moreover, my gratitude goes to Professor Jeffrey Cohn at University of Pittsburgh for hosting me in his lab and meeting his team, who were impressing with the innovative knowledge every member have.

It should not be odd to thank the Australian people in general and the Canberra people in particular for making living away from home almost like being home. Their friendly, humble and accepting nature, as well as the smiles, and greetings made me feel welcome and reduced the homesickness.

The support of the administrative staff at the College of Engineering and Computer Science at the Australian National University is much appreciated. I thank them for their patience and assistance with my continuous pushing to finalise my paperwork, starting from the PhD offer to submitting my thesis.

I wish to thank all my friends and colleagues, for being there whenever I needed a social life and being understanding when I disappeared diving into my work.

By this point, I hope that I have already expressed my thanks to my husband and family, even though I would never be able to thank them enough for their prayers, and for their waiting and emotions every time I arrive and leave my home country during these years.

Finally, I acknowledge and appreciate the receipt of the King Abdullah scholarship, which has been very generous and the staff at the Saudi Arabia Culture Mission for their supportive actions.

Abstract

Clinical depression is a common mood disorder that may last for long periods, vary in severity, and could impair an individual's ability to cope with daily life. Depression affects 350 million people worldwide and is therefore considered a burden not only on a personal and social level, but also on an economic one. Depression is the fourth most significant cause of suffering and disability worldwide and it is predicted to be the leading cause in 2020.

Although treatment of depression disorders has proven to be effective in most cases, misdiagnosing depressed patients is a common barrier. Not only because depression manifests itself in different ways, but also because clinical interviews and self-reported history are currently the only ways of diagnosis, which risks a range of subjective biases either from the patient report or the clinical judgment. While automatic affective state recognition has become an active research area in the past decade, methods for mood disorder detection, such as depression, are still in their infancy. Using the advancements of affective sensing techniques, the long-term goal is to develop an objective multimodal system that supports clinicians during the diagnosis and monitoring of clinical depression.

This dissertation aims to investigate the most promising characteristics of depression that can be "heard" and "seen" by a computer system for the task of detecting depression objectively. Using audio-video recordings of a clinically validated Australian depression dataset, several experiments are conducted to characterise depression-related patterns from verbal and nonverbal cues. Of particular interest in this dissertation is the exploration of speech style, speech prosody, eye activity, and head pose modalities. Statistical analysis and automatic classification of extracted cues are investigated. In addition, multimodal fusion methods of these modalities are examined to increase the accuracy and confidence level of detecting depression. These investigations result in a proposed system that detects depression in a binary manner (e.g. depressed vs. non-depressed) using temporal depression behavioural cues.

The proposed system: (1) uses audio-video recordings to investigate verbal and nonverbal modalities, (2) extracts functional features from verbal and nonverbal modalities over the entire subjects' segments, (3) pre- and post-normalises the extracted features, (4) selects features using the T-test, (5) classifies depression in a binary manner (i.e. severely depressed vs. healthy controls), and finally (6) fuses the individual modalities.

The proposed system was validated for scalability and usability using generalisation experiments. Close studies were made of American and German depression datasets individually, and then also in combination with the Australian one. Applying the proposed system to the three datasets showed remarkably high classification

results - up to a 95% average recall for the individual sets and 86% for the three combined. Strong implications are that the proposed system has the ability to generalise to different datasets recorded under quite different conditions such as collection procedure and task, depression diagnosis testing and scale, as well as cultural and language background. High performance was found consistently in speech prosody and eye activity in both individual and combined datasets, with head pose features a little less remarkable. Strong indications are that the extracted features are robust to large variations in recording conditions. Furthermore, once the modalities were combined, the classification results improved substantially. Therefore, the modalities are shown both to correlate and complement each other, working in tandem as an innovative system for diagnoses of depression across large variations of population and procedure.

List of Publications

Thesis related:

- S. Alghowinem, R. Goecke, J. Cohn, M. Wagner, G. Parker and M. Breakspear. Cross-Cultural Detection of Depression from Nonverbal Behaviour. Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG 2015), Ljubljana, Slovenia, 4-8 May 2015. (Accepted 13 Jan 2015, presented 6 May 2015)

- S. Alghowinem, R. Goecke, M. Wagner, G. Parker and M. Breakspear. Eye Movement Analysis for Depression Detection. Proceedings of the 2013 IEEE International Conference on Image Processing ICIP2013, pages 4220-4224, Melbourne, Australia, 15-18 Sep 2013.

- S. Alghowinem, R. Goecke, M. Wagner, G. Parker and M. Breakspear. Head Pose and Movement Analysis as an Indicator of Depression. Proceedings of the Fifth Biannual Humaine Association Conference on Affective Computing and Intelligent Interaction ACII2013, pages 283-288, Geneva, Switzerland, 2-5 Sep 2013.

- S. Alghowinem. From Joyous to Clinically Depressed: Mood detection using multi-modal analysis of a person's appearance and speech. Doctoral Consortium paper. Proceedings of the Fifth Biannual Humaine Association Conference on Affective Computing and Intelligent Interaction ACII2013, pages 648-654, Geneva, Switzerland, 2-5 Sep 2013.

- S. Alghowinem, R. Goecke, M. Wagner, J. Epps, G. Parker and M. Breakspear. Characterising Depressed Speech for Classification. Proceedings of the 14th Annual Conference of the International Speech Communication Association INTERSPEECH2013, pages 2534-2538, Lyon, France, 25-29 Aug 2013.

- J. Joshi, R. Goecke, S. Alghowinem, A. Dhall, M. Wagner, J. Epps, G. Parker and M. Breakspear. Multimodal Assistive Technologies for Depression Diagnosis and Monitoring. Journal on Multimodal User Interfaces, 7(3): 217-228, Springer, 2013.

- S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Breakspear and G. Parker. Detecting Depression: A Comparison between Spontaneous and Read Speech. Proceedings of the 38th International Conference on Acoustics, Speech, and Signal Processing ICASSP2013, pages 7547-7551, Vancouver, Canada, 26-31 May 2013.

- S. Alghowinem, R. Goecke, M. Wagner, J. Epps, T. Gedeon, M. Breakspear and G. Parker. A Comparative Study of Different Classifiers for Detecting Depression from Spontaneous Speech. Proceedings of the 38th International Conference on Acoustics, Speech, and Signal Processing ICASSP2013, pages 8022-8026, Vancouver, Canada, 26-31 May 2013.
- S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Breakspear, and G. Parker. From Joyous to Clinically Depressed: Mood Detection Using Spontaneous Speech. Proceedings of the 25th International FLAIRS Conference FLAIRS-25, pages 141-146, Marco Island (FL), USA, 23-25 May 2012.

Other publications during PhD training (not directly related to this thesis):

- M. AlShehri, S. Alghowinem. An exploratory study of detecting emotion states using pupil size and fixation with the eye-tracking technology. Proceedings of the Science and Information Conference (SAI 2013), pages 428-433, London, UK, October 7-9, 2013.
- S. Alghowinem, S. Alghuwinem, M. Alshehri, A. Alwabil, R. Goecke and M. Wagner. Design of an Emotion Elicitation Framework for Arabic Speakers. Proceedings of the 16th International Conference on Human-Computer Interaction (HCI International 2014), pages 717-728, Heraklion, Greece, 22-27 June 2014.
- S. Alghowinem, M. AlShehri, R. Goecke and M. Wagner. Exploring Eye Activity as an Indication of Emotional States Using an Eye-tracking Sensor. In L. Chen, S. Kapoor and R. Bhatia (Eds.), Intelligent Systems for Science and Information: Extended and Selected Results from the Science and Information Conference 2013, Studies in Computational Intelligence, Vol. 542, pages 261-276, Springer, 2014.
- S. Alghowinem, M. Wagner and R. Goecke. AusTalk, The Australian Speech Database: Design Framework, Recording Experience and Localisation. Proceedings of the 8th International Conference on IT in Asia 2013 (CITA'13), pages 1-7, Malaysia, 1-4 July 2013.

Contents

Acknowledgments vii

Abstract ix

List of Publications xi

Glossary xxi

List of Abbreviations xxiii

1 Introduction 1
- 1.1 Motivation . . . 1
- 1.2 Aim . . . 2
- 1.3 Objectives . . . 3
- 1.4 Research Questions . . . 3
- 1.5 Thesis Outline . . . 4

2 Depression Analysis: A Literature Review 7
- 2.1 Depression . . . 7
  - 2.1.1 Definition and Diagnosis . . . 7
  - 2.1.2 Depression Symptoms . . . 8
    - 2.1.2.1 Verbal Depression Symptoms . . . 9
    - 2.1.2.2 Nonverbal Depression Symptoms . . . 12
- 2.2 Approaches for Sensing Affect . . . 14
  - 2.2.1 General Affect Datasets Collection . . . 15
  - 2.2.2 Data Preprocessing . . . 17
  - 2.2.3 General Feature Extraction . . . 18
    - 2.2.3.1 Audio Feature Extraction . . . 18
    - 2.2.3.2 Video Feature Extraction . . . 20
  - 2.2.4 Features for Emotion Recognition . . . 23
  - 2.2.5 Feature Selection . . . 24
  - 2.2.6 Classification . . . 26
  - 2.2.7 Multimodal Affective Sensing (Fusion Approaches) . . . 28
  - 2.2.8 Evaluation and Validation of Affective Systems . . . 30
- 2.3 Depression Recognition Using Computer Techniques . . . 32
  - 2.3.1 Depression Datasets . . . 33
  - 2.3.2 Detecting Depressed Speech . . . 37
  - 2.3.3 Detecting Nonverbal Depressed Behavior . . . 39

2.3.4 Multimodal Depression Detection . . . 40
2.3.5 Cross-cultural Depression Detection . . . 41
2.4 Summary . . . 45

3 Overall System Design and Datasets 47
3.1 System Design, Analysis and Evaluation . . . 47
3.2 Datasets . . . 57
3.2.1 BlackDog Dataset . . . 57
3.2.2 Pitt Depression Dataset . . . 61
3.2.3 AVEC German Depression Dataset . . . 62
3.3 Summary . . . 63

4 Depressed Speech Characteristics 67
4.1 Speech Signal Pre-processing . . . 67
4.2 Speaking Style . . . 69
4.2.1 Extracting Speaking Style Features . . . 70
4.2.2 Statistical Analysis of Speaking Style Features . . . 71
4.2.3 Classification Using Speaking Style Features . . . 73
4.3 Vocal Prosody . . . 77
4.3.1 Extracting Vocal Prosody Features . . . 77
4.3.2 Statistical Analysis of Vocal Prosody Features . . . 79
4.3.3 Classification Using Vocal Prosody Features . . . 81
4.4 Summary . . . 86

5 Depressed Appearance Characteristics 89
5.1 Eye Activity . . . 89
5.1.1 Locating and Tracking the Eyes . . . 90
5.1.2 Eye Activity Feature Extraction . . . 92
5.1.3 Statistical Analysis of Eye Features . . . 94
5.1.4 Classification Using Eye Activity Features . . . 96
5.2 Head Pose and Movement . . . 100
5.2.1 Locating the Face to Estimate Head Pose . . . 101
5.2.2 Head Pose and Movement Feature Extraction . . . 103
5.2.3 Statistical Analysis of Head Features . . . 104
5.2.4 Classification Using Head Pose Features . . . 105
5.3 Summary . . . 109

6 Multimodal Fusion 111
6.1 Fusion Methods . . . 111
6.2 Fusion Results . . . 115
6.3 Classifier Error Analysis . . . 118
6.4 Classifier Comparison and Fusion . . . 120
6.4.1 Classifier Comparison . . . 120
6.4.2 Classifier Fusion . . . 122
6.5 Summary . . . 124

7 Generalisation Across Datasets **127**
7.1 Proposed System . . . 127
7.1.1 Description of this proposed system applied on BlackDog dataset (a review) . . . 127
7.1.2 Differences of Applying the Concluded System on AVEC and Pitt datasets . . . 129
7.2 Results of Generalisation on Individual Datasets . . . 132
7.3 Results of Generalisation on Combinations of Datasets . . . 134
7.4 Summary . . . 143

8 Conclusions **147**
8.1 Research Questions . . . 147
8.2 Summary of Contributions . . . 149
8.3 Future Work . . . 151

Appendix A: Classification Results for each Subject for each Modality. **155**

List of Figures

2.1 General verbal depression symptoms . . . 10
2.2 General nonverbal depression symptoms . . . 12
2.3 Stages to be considered in a multimodal emotion recognition systems . 15
2.4 General audio feature categories . . . 18
2.5 General video feature extraction method . . . 21
2.6 General feature selection approaches . . . 25
2.7 General pattern recognition categories . . . 26
2.8 General fusion approaches . . . 28

3.1 Overview of system design and selected method . . . 48
3.2 A view of the BlackDog recording environment setup . . . 60
3.3 A view of the Pitt recording environment setup [Yang et al., 2013] . . . 61

4.1 An example of manual labelling of the BlackDog dataset interview part 68
4.2 An example of applying the voice activity detector on a subject segment 69

5.1 Final eye AAM model (for open eye state) with 74 points in the order shown . . . 91
5.2 Eye AAM with 74 points for both eyes including eyebrows . . . 92
5.3 Extracting and normalising eye movement features . . . 93
5.4 Extracting and normalising eyelid distance and blink features . . . 93
5.5 Head rotation angles: Yaw, pitch and roll . . . 101
5.6 3D to 2D AAM projection for estimating the head pose . . . 102

6.1 Summary of the feature preparation, extraction, and selection . . . 112
6.2 Feature fusion method . . . 113
6.3 Decision fusion method . . . 113
6.4 Score fusion method . . . 114
6.5 Hybrid fusion approaches . . . 115
6.6 Constructing fuzzy signature . . . 121
6.7 Classifier fusion in the classifiers level approach *(percentages between parentheses are the accuracy classification results in terms of AR from each modality and from each fusion level)* . . . 123
6.8 Classifier fusion in the modalities level approach *(percentages between parentheses are the accuracy classification results in terms of AR from each modality and from each fusion level)* . . . 124

7.1 Average recall of classification results of individual datasets 133

List of Tables

2.1 Confusion matrix for a 2-class problem 31
2.2 Comparison of datasets used for automatic depression detection 34
2.3 Summary of automatic depression recognition studies 43

3.1 Summary of datasets specification used in this research 58

4.1 Gender-independent significant T-test results of speaking style features for the interview 71
4.2 Correct classification results (in %) of speech style features 73
4.3 Total duration of the several segments of the interviews (in minutes) for each gender in each group, as well as average duration and standard deviation (in minutes) 74
4.4 Gender-dependent correct classification results (in %) using speech style features 75
4.5 Total duration of speech of positive and negative questions from the interview (in minutes) for each group, as well as average duration and standard deviation (in minutes) 76
4.6 Expression-dependent correct classification results (in %) using speech style features 76
4.7 Gender-independent significant T-statistic results of speaking prosody features for the interview 80
4.8 Correct classification results (in %) of individual speech prosody features 82
4.9 Correct classification results (in %) of speech prosody features 83
4.10 Total duration of the subjects' sounding segments of the interviews (in minutes) for each gender in each group, as well as average duration and standard deviation (in minutes) *(extract from Table 4.3)* 84
4.11 Gender-dependent correct classification results (in %) using speech prosody features 84
4.12 Total duration of subjects' sounding of positive and negative questions from the interview (in minutes) for each group, as well as average duration and standard deviation (in minutes) *(extract from Table 4.5)* . . 85
4.13 Expression-dependent correct classification results (in %) using speech prosody features 85

5.1 Gender-independent significant T-test results of eye movement features for the interview 95

5.2 Correct classification results (in %) of eye activity low-level and functional features . . . 96
5.3 Total duration of the entire interview (in minutes) for each gender in each group, as well as average duration and standard deviation (in minutes) . . . 98
5.4 Gender-dependent correct classification results (in %) using eye activity functional features . . . 98
5.5 Total duration of positive and negative questions from the interview (in minutes) for each group, as well as average duration and standard deviation (in minutes) . . . 99
5.6 Expression-dependent correct classification results (in %) using eye activity functional features . . . 100
5.7 Gender-independent significant T-test results of head pose and movement features . . . 104
5.8 Correct classification results (in %) of head pose low-level and functional features . . . 106
5.9 Gender-dependent correct classification results (in %) using head pose functional features . . . 107
5.10 Expression-dependent correct classification results (in %) using head pose functional features . . . 108

6.1 Classification results for fused modalities using different fusion methods 116
6.2 Number of misclassified subjects in each modality . . . 118
6.3 Correct classification results (in %) when using different classifiers . . . 121

7.1 Overview of selected methods of the investigations performed on the BlackDog dataset . . . 128
7.2 Classification results of individual datasets . . . 132
7.3 List of fixed features that exceed the T-statistic for the majority of dataset combinations . . . 136
7.4 Classification results of dataset combinations using leave-one-out cross-validation . . . 138
7.5 Classification results and number of selected features of dataset combinations using train-test method . . . 140

Glossary

The same terminology could be used to mean different things based on the field of study. This section provides the intending meaning of some words used in this thesis.

Depression: refereed to different types of clinical depression mental disorder (either Melancholia, bipolar or Major Depression Disorder (MDD)).

Channel: refereed to different signal input (e.g. audio signal, video signal, brain signal, etc.). Channels used in this work are audio and video channels.

Modality: refereed to a specific part (area) of a channel. For example, speech prosody modality is part of the audio channel, where several speech prosody features are extracted. Modalities used in this work are: speech style, speech prosody, eye activity, and head pose modalities.

List of Abbreviations

- **AAM** Active Appearance Model
- **ACF** Auto-Correlation Function
- **AKN** Acknowledgment
- **ANN** Artificial Neural Network
- **ANOVA** Analysis of Variance
- **AR** Average Recall
- **ASM** Active Shape Model
- **AU** Action Unit
- **AVDLC** Audio-Video Depressive Language Corpus
- **AVEC** Audio/Visual Emotion Challenge Depression Dataset
- **BDI** Beck Depression Inventory
- **BlackDog** The Black Dog Institute Depression Dataset
- **BO** Overlap Speech
- **C** Control Group
- **CBT** Cognitive-Behavioral Therapy
- **CESD** Center for Epidemiologic Studies Depression Scale
- **CFS** Correlation-based Feature Selection
- **CLM** Constrained Local Models
- **D** Depressed Group
- **DAIC** The Virtual Human Distress Assessment Interview Corpus
- **DET** Error Trade Off
- **DOF** Degree of Freedom
- **DSM-IV** Diagnostic and Statistical Manual of Mental Disorders

- **DSS** Dynamic Score Selection
- **ECT** Electroconvulsive Therapy
- **EEG** Electroencephalography
- **EER** Equal Error Rate
- **EOH** Edge Orientation Histogram
- **ETF** Exceeded T-statistic Features
- **F0** Fundamental Frequency
- **FACS** Facial Action Coding System
- **FCM** Fuzzy C Mean
- **FN** False Negatives
- **FP** False Positives
- **fps** Frames per Second
- **GA** Genetic Algorithms
- **GMM** Gaussian Mixture Model
- **HFS** Hierarchical Fuzzy Signature
- **HMM** Hidden Markov Models
- **HNR** Harmonic-to-Noise Ratio
- **HRSD** Hamilton Rating Scale for Depression
- **HTK** Hidden Markov Model Toolkit
- **IAPS** International Affective Picture System
- **ICA** Independent Component Analysis
- **ICD-9-CM** International Classification of Diseases, Ninth Edition, Clinical Modification
- **IG** Information Gain
- **LBP** Local Binary Pattern
- **LDA** Linear Discriminant Analysis
- **LFCC** Linear-Frequency Cepstral Coefficients
- **LLD** Low-Level Descriptors

- **log** Logarithmic
- **LOO** Leave-One-Out
- **MDD** Major (clinical) Depressive Disorder
- **MDS** Multidimensional Scaling
- **MFCC** Mel-Frequency Cepstral Coefficients
- **MHH** Motion History Histograms
- **MLP** Multilayer Perceptron Neural Network
- **NN** Nearest Neighbour
- **NPV** Negative Predictive Value
- **ORI** The Oregon Research Center in USA Dataset
- **ORYGEN** The Youth Health Research Centre in Melbourne, Australia Dataset
- **PC** Principal Component
- **PCA** Principle Component Analysis
- **PDBH** The Psychiatry Department of Behavioural Health Dataset at the Medical College of Georgia in the USA
- **PHQ-9** Patient Health Questionnaire-Depression
- **Pitt** The University of Pittsburgh Depression Dataset
- **PLS** Partial Least Squares
- **POSIT** Pose from Orthography and Scaling with ITerations
- **QDA** Quadratic Discriminant Analysis
- **QIDS-SR** Quick Inventory of Depressive Symptoms-Self Report
- **RA** Research Assistant
- **RBF** Radial Basis Function
- **RMS** Root Mean Square
- **RMSE** Root Mean Square Error
- **ROC** Receiver Operating Characteristic
- **ROI** Region of Interest
- **SARS** Severe Acute Respiratory Syndrome

- **SB** Subject Speech
- **SDC** Suicidal, Depressed, and Control Subjects Dataset
- **SIL1** First Silence Lag
- **SIL2** Second Silence Lag
- **STIP** Space Time Interest Points
- **SVM** Support Vector Machine
- **SVR** Support Vector Regression
- **TAT** Thematic Apperception Test
- **TEO** Teager Energy Operator
- **TN** True Negatives
- **TP** True Positives
- **UBM** Universal Background Model
- **VAD** Voice Activity Detection
- **WCGS** The Western Collaborative Group Study Speech Dataset
- **WHO** World Health Organization
- **WTAR** Wechsler Test of Adult Reading

Chapter 1

Introduction

I believe that recent developments in affective sensing technology will potentially enable an objective assessment of mood disorders. Applying and investigating such techniques and overcoming their limitations are the goal of the research described in this thesis. These investigations are carried out on the example of developing affective sensing technology that can assist clinicians in the task of diagnosing depression more objectively and accurately.

1.1 Motivation

Changes in affective state are a normal characteristic of human beings. However, when these changes increase in intensity, last longer, and impact negatively on a person's functioning, a clinical depression line might be crossed. Unlike emotions, which are short term, mood is a long term affective state. Therefore, clinical depression is a mood disorder that may last for weeks, months, even years, vary in severity, and could result in unbearable pain if appropriate treatment is not received. Clinical (or major) depression is different from feeling depressed; it is generally acknowledged to be more serious, last for long periods and affect a person's functioning. The World Health Organization (WHO) lists depression as the fourth most significant cause of suffering and disability world wide and predicts it to be the leading cause in 2020 [World Health Organization, 2003; Mathers et al., 2008]. A recent WHO report estimated that 350 million people worldwide are affected by depression [World Health Organization, 2012]. At its most severe, depression is associated with half of all suicides and presents a significant annual national economic burden [Lecrubier, 2000; US Department of Health and Human Services, 2000]. Moreover, the suicide risk is more than 30 times higher among depressed patients than among the general population [Guze and Robins, 1970].

The statistics are consistent, if not higher, in Australia compared to the rest of the world. The Australian Survey of Mental Health and Well-Being (1997) reported that 6.3% of the population suffer from clinical depression in any one year, noting that this percentage does not include people who choose not to get professional help. From an economic point of view, depression is a national economic burden as six million working days are lost each year to depression and ten million antidepres-

sant prescriptions are written every year. More than 180 Australians take their life in depression related suicide each month [Australian Bureau of Statistics, ABS, 2008]. Even though people of all ages suffer from depression, Australia has one of the highest depression related youth suicide rates compared to several countries around the world [Prendergast, 2006]. According to Prendergast [2006], this can be prevented if depressed subjects seek help from professionals, and if health professionals could be provided with suitable objective technology for detecting and diagnosing depression. Therefore, recognising depression in primary care is a critical public health problem [Baik et al., 2005].

Although several treatment methods of depression disorders exist and are effective in most cases [Kiloh et al., 1988; Simon et al., 1998; Pence et al., 2012], misdiagnosing depressed patients is a common barrier [Niedermaier et al., 2004]. Based on the WHO Global Burden of Disease report [World Health Organization, 2012], the barriers to effective diagnosis of depression include a lack of resources and trained health care providers. Even though clinical depression is one of the most common mental disorders, it is often difficult to diagnose. Not only because it manifests itself in different ways, but also because the assessment methods for diagnosing depression rely almost exclusively on patient-reported or clinical judgments of symptom severity [Albrecht and Herrick, 2010; Mundt et al., 2007], risking a range of subjective biases either from the patient report or the clinical judgment. Moreover, evaluations by clinicians vary depending on their expertise and the diagnostic methods used (e.g. Diagnostic and Statistical Manual of Mental Disorders (DSM-IV)American Psychiatric Association [1994], Quick Inventory of Depressive Symptoms-Self Report (QIDS-SR) Rush et al. [2003], Hamilton Rating Scale for Depression (HRSD)Hamilton [1960], Beck Depression Inventory (BDI)Beck et al. [1996], etc.). Currently, there is no objective method to diagnose depression, nor a laboratory-based test for diagnosing depression. Rather, it is diagnosed as part of a complete mental health evaluation.

The motivation behind the research described in this dissertation is to investigate how affective sensing technology can play a role in providing an objective assessment of depression.

1.2 Aim

The goal of this research is to develop an objective affective sensing system that supports clinicians in their diagnosis of clinical depression and its level from a person's visual appearance and speech. This system could progress towards a diagnostic aid, which will be clinically tested in collaboration with the Black Dog Institute (Sydney, Australia). This system might be used for supporting the computer based Cognitive-Behavioral Therapy (CBT) with an ability to detect improvement in a patient's mood. In the long term, such an objective multimodal affective sensing system may also become a very useful tool for remote depression monitoring to be used for doctor-patient communication in the context of an e-health infrastructure. In addition, since video and audio channels are complementary rather than redundant, fusing multi-

modal cues will improve depression detection [Cohn et al., 2009]. Therefore, in order to get a more accurate diagnosis, the system will use multimodal analysis; that is, combinations of speech prosody and behaviour, eye activity and blink, as well as head pose and movement. As facial expressions to diagnose depression have been investigated in the previous literature, and as the video recording does not include full body to analyse body and hand movements, these modalities will not be included in this thesis, but are acknowledged as potential further relevant sources of information.

1.3 Objectives

Of particular interest in this dissertation is analysing the behavioural patterns, which could be detected by a computer system that differentiate depressed patients from healthy controls in a clinical interview context. Comparing depressed subjects with healthy controls, this thesis' major objectives are as follows:

- Explore features extracted from speaker characteristics and visually observed behaviour for their ability to distinguish depressed subjects from healthy control subjects.
- Investigate differences in overall movement patterns of the eyes and the head between depressed and healthy control subjects.
- Examine fusion techniques for speech, eye, and head behavioural cues to increase the depression recognition rate and to improve the robustness of the recognition.
- Generalise and validate research results by applying the investigated methods of feature extraction, feature selection, classification and fusion on different databases of subjects of different cultural backgrounds and languages.

1.4 Research Questions

As stated above, a key motivation for undertaking the research documented in this thesis is a need to develop an affective sensing technology that can assist clinicians in the task of diagnosing depression more accurately. Following the research objectives presented above, the following questions are posed that are investigated here:

Q1. What are the distinguishing characteristics and the most accurate configurations of verbal based depression detection in terms of extracted features, classification methods, and gender-dependence?

Q2. What are the specific nonverbal behaviour, movement, or activity patterns of the eyes and the head that could distinguish depressed patients from healthy control subjects in the classification task of detecting depression?

Q3. Which fusion method of the examined modalities (speech, eye, and head behavioural patterns) would improve the robustness and increase the accuracy of the depression recognition?

Q4. Are the findings and methods data-specific, or would they generalise to give similar results when used on different datasets of different recording environments as well as different cultures and languages?

1.5 Thesis Outline

This dissertation comprises eight chapters, including this introduction. A brief outline of the remaining chapters is as follows:

Chapter 2: This chapter provides a literature review of depression diagnosis and symptoms. Also, it gives a general review of emotion recognition systems and the required steps to build them. The chapter concludes with a deeper review of studies that investigated depression detection using affective sensing technology.

Chapter 3: This chapter describes the general methodology that has been used in this research to build a depression recognition system. The chapter presents the general methods used to prepare, extract, and select the features used for classification. Fusion techniques for the fusion experiment and normalisation for the generalisation experiment are also explained in this chapter. It also includes a full description of the main depression dataset used in this research, as well as descriptions of two other depression datasets that have been used for generalisation.

Chapter 4: Several experiments to recognise depression from speech prosody and behaviour features are presented, which provide an understanding of depressed speech characteristics.

Chapter 5: In this chapter, two experiments conducted on eye activity and head movement are described. This chapter also shows the most distinguishing features from the eyes and head behavioural patterns of depression and their classification results.

Chapter 6: Several techniques for fusing different channels exist. This chapter investigates techniques to fuse all features from speech, eyes, and head behavioural patterns to indicate which fusion technique would give the most accurate result.

Chapter 7: In this chapter, an experiment is conducted to validate and generalise the findings of detecting depression. That is done by applying the investigated methods on different datasets with different recording environments, cultural backgrounds and languages.

Chapter 8: Finally, the conclusions and a summary of the contributions of this dissertation are presented. Open issues and future directions for ongoing research are discussed.

Chapter 2

Depression Analysis: A Literature Review

Understanding depression and its symptoms is a first step for developing a system that could recognise and diagnose it. In this chapter, a background review of clinical depression and its symptoms is given. Moreover, a general review of automatic emotion recognition systems, and the required steps to build them are described, which will help understand the requirements for building a depression recognition system. Finally, this chapter concludes with a deeper review of studies that investigated depression detection using affective sensing technology.

2.1 Depression

2.1.1 Definition and Diagnosis

Clinical depression is a serious illness that affects not only mental, but also physical health. Albrecht and Herrick [2010] defines clinical depression as a medical condition that affects and changes a person's thoughts, mood, and behaviours as well as the body. This condition is associated with various physical problems, such as fluctuation of sleep patterns, appetite, and energy. The physical symptoms improve with depression treatment. Without treating depression, these physical symptoms such as loss of energy, make a person's functioning harder. It has been argued that depression is not a result of personal or moral weakness but is a treatable illness [Albrecht and Herrick, 2010]. Clinical (or major) depression is different from feeling depressed; it is generally acknowledged to be more serious, lasts for long periods and affects a person's functioning. It has been agreed that depression might be caused by genetic factors, medications, personality, life events, drugs and/or alcohol [Albrecht and Herrick, 2010]. Different types of depression exist and can be distinguished based on the symptoms and diagnosing method. Common types of depression are: major (clinical) depressive disorder (MDD), dysthymic (chronic) depression disorder, bipolar (manic) depression, postpartum depression, seasonal affective disorder, etc. Prendergast [2006].

Depression has no dedicated laboratory tests or procedures for it to be diag-

nosed; it is diagnosed as part of a complete mental health evaluation. It depends on a person's symptoms self-report and a professional observation. The evaluation includes current and past symptoms, family history, and medical history as well as mental status [Albrecht and Herrick, 2010]. However, professionals' evaluations vary depending on their expertise and the diagnosing methods being used. Several diagnostic tests exist (e.g Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) [American Psychiatric Association, 1994], Hamilton Rating Scale for Depression (HRSD) [Hamilton, 1960], Beck Depression Inventory (BDI) [Beck et al., 1996], Patient Health Questionnaire-Depression (PHQ-9) [Kroenke and Spitzer, 2002], etc.). Those tests vary in number of items and scoring points, which leads to inconsistencies in diagnosing depression between psychiatrists. Moreover, even when only one test is used, subjective views and experiences of psychiatrists could result in differences on the diagnosis. That is, two psychiatrists using the same diagnosis test are more likely to come with a different score for the same subject. As a result, there currently is no objective method to diagnose depression.

2.1.2 Depression Symptoms

The American Psychiatric Association outlined criteria for mental disorders including depression symptoms (DSM-IV) [American Psychiatric Association, 1994]. For a positive diagnosis of depression, five or more of DSM-IV criteria have to be present for at least two weeks. In addition, other symptoms might be derived from DSM-IV such as: sadness or irritability, and unexplained physical complaints (e.g. headache, backache, stomach upset) [Albrecht and Herrick, 2010]. Some of these symptoms include:

- loss of interest or pleasure in activities,
- feelings of worthlessness or inappropriate guilt,
- psychomotor agitation or retardation,
- fatigue or loss of energy, and
- recurrent thoughts of death or suicidal ideation.

DSM-IV classifies depression as a mood disorder that implicates deficient positive affect, excessive negative affect, or both. [Ekman and Fridlund, 1987] and Ekman [1994] confirmed that negative affect is dominant, had longer duration, and had higher intensity in depressed patients than control subjects.Moreover, it is believed that depression influences emotional reactions, where several studies were conducted to study this influence, as reviewed in Bylsma et al. [2008]. The review study revealed that depression was characterised by reduced emotional activity to both negative and positive valenced stimuli, with the reduction larger for positive stimuli [Bylsma et al., 2008]. Moreover, cultural acceptance and recognition of depression disorder affect the way depressed patients report their symptoms. Noting differences in depression prevalence between countries, cultural psychiatrists were

keen to investigate the reasons behind these differences. Yet underdeveloped, initial investigations found that depression is manifested in different symptoms depending on the culture. In particular, somatic or cognitive symptoms are displayed differently based on cultural background [Singer, 1975; Tseng, 2001; Ruchkin et al., 2006]. Singer [1975] reviewed cultural studies of depression in term of prevalence and manifestation of depression illness of several countries. According to the review, depression in western cultures display itself commonly as cognitive symptoms such as psychomotor retardation and guilt, while in non-western cultures, such as from Africa and India, depressed patients report somatic symptoms such as stomachache, backache or sexual dysfunction [Singer, 1975; Tsai and Chentsova-Dutton, 2002]. Three western countries (United States, Belgium, and Russia) have been compared in the Ruchkin et al. [2006] study, suggesting that depression patterns are similar in those populations. Such cultural differences in depression manifestation lead to differences in depression diagnosis, which make comparing depression diagnosis in different countries a difficult task [Singer, 1975; Tseng, 2001]. Comparing difficulty is not only because each country uses different diagnostic methods and scales, but also the differences in reported symptoms, which make it difficult to find an objective method of diagnosing depression in cross-cultural context. Therefore, this work will investigate depression symptoms from Australian, American, and German subjects, hypothesising to find objective depression symptoms of similar cultures. Future work could investigate finding objective depression symptoms of subjects of different cultures.

However, to give a computer the ability to detect depression using verbal and nonverbal cues, only symptoms of depression that are interpretative to the computer sensor devices could be explored. As mentioned earlier, depression is a mood disorder since it lasts longer than emotions. However, there are no distinguished vocal, facial or body expressions for mood; rather, mood could be measured by identifying emotions that are associated with it and then aggregating them over time [Ekman and Davidson, 1994; Ekman et al., 1997a]. Ekman [1999] suggested that identifying these associated emotions, their strength, and their repetitive sequence as well as whether the emotion was spontaneous might help to diagnose and monitor treatment progress of depression.

2.1.2.1 Verbal Depression Symptoms

Studies investigating vocal affect of depressed subjects are more advanced than nonverbal affect. Studies of psychology investigations of depressed speech have found several distinguishable prosody features (see Figure 2.1), such as:

Pitch: which refers to the tonal height of a sound (although inaccurate often referred to as fundamental frequency F0), has been widely investigated in the depression literature. A lower range of pitch indicates a monotone speech, which is a common feature of depressed speech [Nilsonne, 1988; Ellgring and Scherer, 1996; Moore et al., 2004; Mundt et al., 2007; Kuny and Stassen, 1993; Ellgring and Scherer, 1996]. Pitch variance is more emphasised for healthy controls and

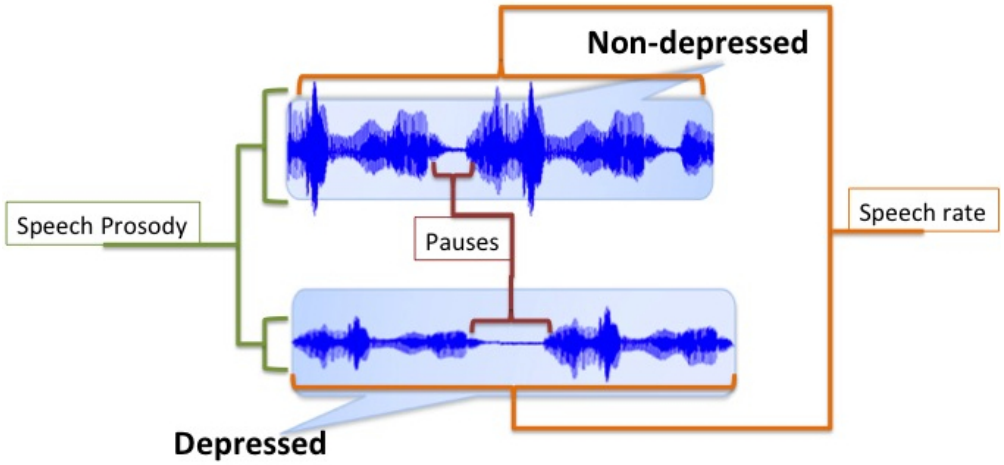


Figure 2.1: General verbal depression symptoms

subdued for depressed subjects, while the pitch range increases after depression treatment [Ozdas et al., 2000]. A lower variance of pitch indicates a lack of normal expression as can be noticed in depressed patients [Moore et al., 2008]. A lower range and variance of pitch is associated to psychomotor retardation, which is one of the main symptoms of depression.

Loudness: refers to the magnitude of sound in a specified direction (also referred to as sound level). There is convincing evidence that sadness and depression are associated with a decrease in loudness [Scherer, 1987], resulting in reduced loudness for depressed subjects.

Energy: Vocal source energy is also a distinguishable feature for depression, resulting in lower energy in the glottal pulses for depressed patients [Ozdas et al., 2004].

Formants: are defined as the spectral peaks of the sound spectrum, which identify vowels. The formant with the lowest frequency is called F1, the second F2, and so on. Formants are a widely used feature in the affect literature [Koolagudi and Rao, 2012; Flint et al., 1993; Moore et al., 2008], being a significantly distinguishable feature for depression [Flint et al., 1993; Moore et al., 2008]. That is due to the fact that psychomotor retardation as a symptom of depression can lead to a tightening of the vocal tract, which tends to affect the formant frequencies [France et al., 2000]. Moreover, of the first three formants [Flint et al., 1993; Moore et al., 2008], a noticeable decrease in the second formant frequency was shown for depressed individuals compared to controls [Flint et al., 1993].

Jitter: measures frequency variability in the pitch of the vocal note in comparison to the fundamental frequency. Higher jitter levels suggest that something is interfering with normal vocal fold vibration. Jitter voice feature have been

analysed, finding higher jitter in depression caused by the irregularity of the vocal fold vibrations [Scherer, 1987; Nunes et al., 2010].

Shimmer: measures amplitude variability in the amplitude of the vocal note in comparison to the fundamental frequency. Higher shimmer levels can reflect problems in neuromuscular control. Unlike jitter, shimmer is lower for depressed subjects [Scherer, 1987; Nunes et al., 2010].

HNR (Harmonic-to-Noise Ratio) quantifies the relative amount of additive noise in the voice signal. HNR is used to connect physiological aspects of voice production to a perceptual impression of the voice, because the degree of spectral noise is related to the quality of the vocal output [de Krom, 1993]. Like the jitter feature, HNR values are higher for depressed subjects, due to the fact that patterns of air flow in the speech production differ between depressed and control subjects [Low et al., 2011].

In addition to the prosody features above, speech style or behavioural based evaluations of depressed speech have found several distinguishing speech patterns as indicators of disease progression, severity or treatment efficacy. Behavior speech features include, but are not limited to the following:

Speaking rate: measures the relative speed or slowness of utterance. It has been found that depressed patients have a slower rate of speech when compared to normal speaking patterns [Moore et al., 2004, 2008; Ellgring and Scherer, 1996].

Pause: refers to the pauses in an utterance and between utterances. Research on the vocal indicators in depressed subjects found an increase in pause time [Ellgring and Scherer, 1996; Reed, 2005; Sobin and Sackeim, 1997]. In the Zlochower and Cohn [1996] study, the vocal timing in clinically depressed mothers in response to their infants was measured, and found to be longer and more variable in the duration of silences.

Response time: is the latency to respond or interact to a conversation. Research on the latency of vocal responses in depressed subjects has been investigated [Reed, 2005; Sobin and Sackeim, 1997] and an increase in latency responses in depressed subjects has been found. Zlochower and Cohn [1996] confirmed this by measuring the vocal timing in clinically depressed mothers in response to their infants. Moreover, they found that the response delay increases with the severity level of depression.

Articulation rate: measures the number of words compared to the utterance duration. Articulation rate has been found to be lower in depressed patients [Pope et al., 1970; Ellgring and Scherer, 1996; Sobin and Sackeim, 1997], compared to normal speaking patterns.

Speaking duration: measures the duration of actual speech in a period of time or a conversation. Speaking duration has been found to be shortened in depressed subjects [Sobin and Sackeim, 1997], compared to normal speaking patterns.

2.1.2.2 Nonverbal Depression Symptoms

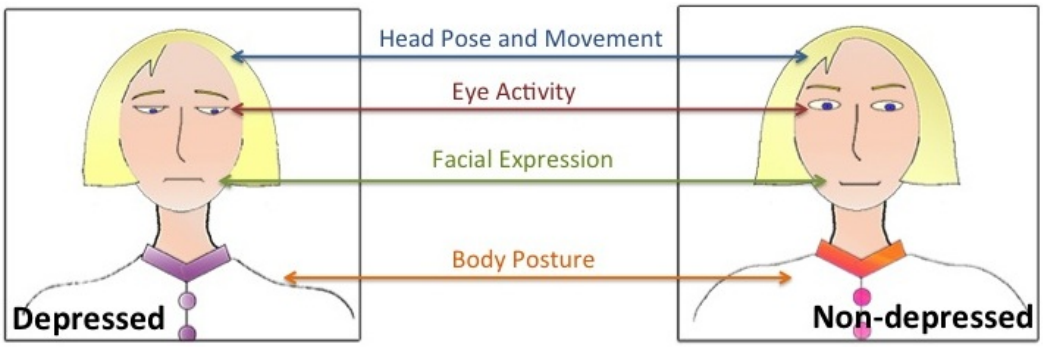


Figure 2.2: General nonverbal depression symptoms

Body language and nonverbal behaviour of depressed subjects have been investigated in psychological studies. Several body language channels have been investigated (see Figure 2.2), such as:

Head pose and movement: Psychological research on depression subjects' head movements has received far less attention compared to the studies on speech, gestures and facial expressions. Simple behaviours such as head movement could reflect cues about mood, emotions, personality, or cognitive processing [Heylen, 2006]. An ethological study on depressed patients' behaviour noticed that behavioural elements are pronounced more in the head and hand regions compared with other body regions [Pedersen et al., 1988]. Fossi et al. [1984] studied the social behaviour of depression finding that depressed patients present significantly less head-nodding than controls. Waxer [1974] found that a depressed person is more likely to position their head downward than a healthy one. Studying eye contact, Fossi et al. [1984] found that depressed patients engage in less eye contact with others than non-depressed persons, illustrated by looking away and a head pose that avoids eye contact. A study investigating depressed patients' involvement in a conversation showed a low involvement reflected by reduced head nodding, fewer head movements, reduced eye contact and gesturing during speech [Hale III et al., 1997].

Eyes activity: If it is true that "eyes are the windows to the soul", gaze is the most illustrative cue in nonverbal communication [Ellgring, 1989]. Research into potential bio-markers of central nervous system disorders such as affective disorders have explored subtle changes in eye movements as possible physiologically based indicators of disease progression, severity or treatment efficacy [Kathmann et al., 2003]. Lipton et al. [1980] found abnormal horizontal pursuit eye movements in depressed patients compared to healthy controls. This was also confirmed in [Abel et al., 1991], finding that pursuit and saccade eye movement rates correlate strongly in controls but are reduced or absent in affective disorder patients, concluding abnormality in patients' ocular motor systems

as a form of psychomotor retardation. Crawford et al. [1995] found the same abnormal saccades in patients even without being under medication.

Furthermore, the results in [Sweeney et al., 1998] identify not only significant eye motor system disturbances in depression but also significant disturbances in their cognitive performance. Nonverbal behaviour, including eye glances and brow raising, hypothesised to have a good potential data source for ascertaining a patient's mental state such as the depression level and to be a clue for effectiveness of treatment [Ekman and Friesen, 1974]. Depressed patients were found to differ from the normal comparison group in decreased direct eye contact with the interviewer and decreased eyebrow movement [Sobin and Sackeim, 1997]. In [Ellgring, 1989], the decrease in emotional and cognitive capacity in depressed subjects was correlated to the reduction and avoidance of gazing at others in a social interaction. Moreover, eye blink rate was investigated showing elevated blink rates, which return to normal levels as the depressed patient's condition improves [Mackintosh et al., 1983].

Facial expression: Several studies investigated facial expressions of depressed subjects. Ekman et al. [1997b] found that depressed subjects show sadness and disgust more often. He also found that depressed subjects have more unfelt smile and less felt smile [Ekman and Fridlund, 1987; Ekman et al., 1997b]. It was confirmed by Ekman [1994] that the dominant emotion in depressed patients is sadness. Moreover, the sadness period in depressed patients last longer and is more intense [Ekman and Fridlund, 1987]. Mouth movement is also an indicator for depression and suicide risk. Nonspeech mouth movement is implicated in subsequent risk for suicide [Cohn et al., 2009]. Lip wiping and wetting could be an indicator of a depressed subject on an antidepressant, as it is a side effect of these medications [Ekman and Fridlund, 1987]. In addition, a recent study concluded that depressed subject lack facial activity [McIntyre et al., 2009].

Body posture and hand movements: In general, the area of recognising the affective state from the body posture is not adequately explored yet. Therefore, few studies on depression sufferers' body posture have been conducted. Depression generally causes a slowing of body movements [Dittmann, 1987], due to a loss of energy and fatigue symptoms. Likewise, depression can result in a slumped body posture [Jackson, 1983], slow and restricted body movements such as stooped or hunched, and tensions demonstrated by rigid posture and movement [France, 2001]. Moreover, depressed subjects engage in more self-body contact (self-touching, including rubbing and scratching) [Jones and Pansa, 1979; Ranelli and Miller, 1981], and significantly less gesturing than healthy controls [Fossi et al., 1984; Ekman and Friesen, 1972].

2.2 Approaches for Sensing Affect

For the past few decades, affective state recognition has been an active research area and has been used in many contexts. Several research areas are involved and contribute to affect recognition, including psychology, speech analysis, computer vision, machine learning and many others. Automatic affective state recognition aims to give computers the ability to observe and interpret affect features. For example, detecting a learner's emotional state could improve the interaction between the learner and computer in a computer-based learning environment [Picard, 1997; Calvo and D'Mello, 2011].

One of the issues faced in developing an automatic affect recognition system is that affect has no clear definition in the psychological literature. Psychologically, affect is a general concept, which covers not only emotions, but also includes mood, attitudes, desires, preferences, intentions, dislikes, etc. [Sloman et al., 2005]. Emotions are distinguished from mood in terms of criteria such as duration, intensity, and association with an inner or outer object [Berrios, 1985]. Emotions are defined as feeling states that are short-lived, more or less intense, and related to a recognisable object [Berrios, 1985]. Mood on the other hand, is defined as longer lasting and objectless states [Berrios, 1985].

Research on distinguishing one affective state from the other is highly controversial, where several methods and emotion models have been proposed. Ekman [1999], for example, studied six basic emotions (happiness, sadness, surprise, fear, anger and disgust) that are recognised universally. However, since emotion could be more complex and blended, alternative ways have been suggested, using multiple dimensions or scales. One way to describe emotions is by a dimensional description including evaluation, activation, control, power, etc. Russell [1979] and Jaimes and Sebe [2007] suggested using a two-dimensional plane with valence and arousal as axes. The valence level represents the quality of the emotion, ranging from unpleasant to pleasant, and the arousal level denotes a quantitative activation level, from not aroused to excited. For example, depression would have low valence and arousal, which places it in the 3rd quarter of the two-dimensional plane. Schlosberg [1954] and Wundt [2009] proposed a three-dimensional emotion model. A few other theories and dimensional models exist. The existence of several models to represent emotions implies a controversial and ambiguous definition of emotions, which leads to difficulties in identifying and labelling emotions for automatic emotion recognition [Gunes et al., 2011]. However, for simplicity, most automatic emotion analysis to date either used discrete emotion classes, such as the basic emotions suggested by Ekman [1999], or the two-dimensional emotion model by Russell [1979].

Regardless of the emotion model used, defining the modality (i.e. verbal, nonverbal) for expressing such emotions is another challenge. Most affective sensing systems use a monomodal recognition system such as voice or face only. Only few systems use multimodal input where different channels are fused, such as audio and video channels, or different modalities such as body movement, facial expression and speech prosody [Pantic and Rothkrantz, 2003; Pantic et al., 2005; Sebe et al.,

2005; Caridakis et al., 2007; Jaimes and Sebe, 2007; D'Mello and Graesser, 2010], as well as fusion with physiological signals [Hussain et al., 2012; Monkaresi et al., 2012]. Based on the aforementioned studies, multimodal systems for recognising emotions are believed to be more accurate than unimodal ones (see Section 2.2.7 for more details). In order to implement multimodal emotion recognition systems, several stages have to be considered as shown in Figure 2.3.

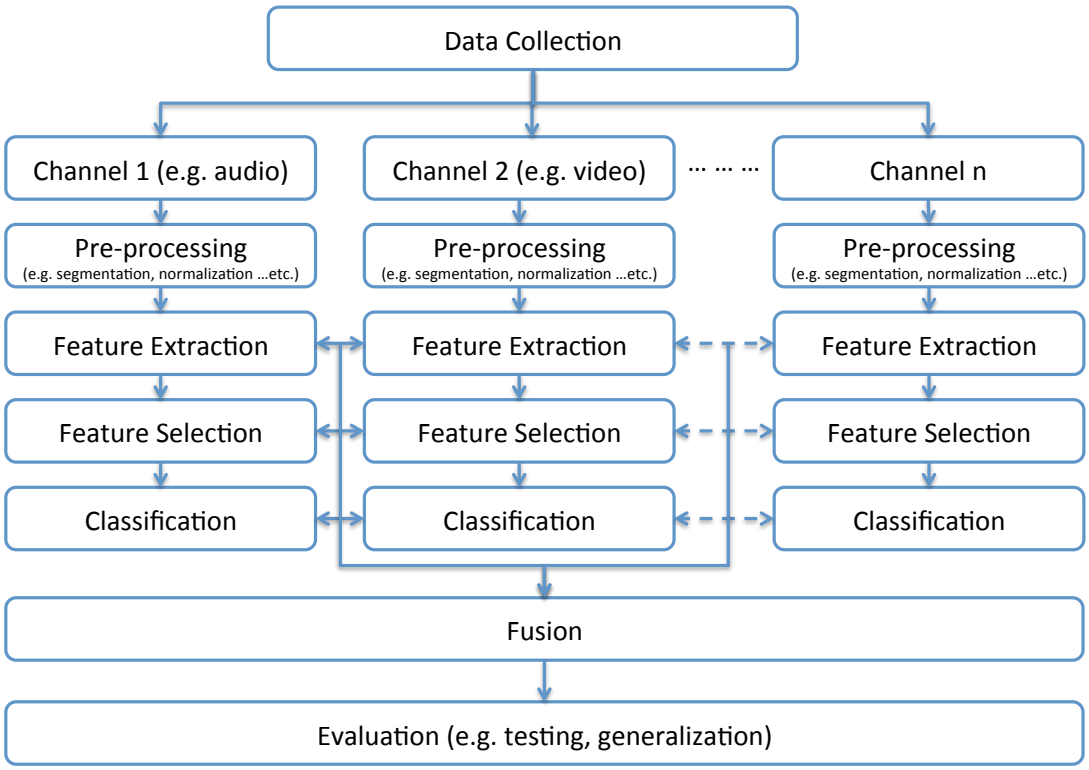


Figure 2.3: Stages to be considered in a multimodal emotion recognition systems

2.2.1 General Affect Datasets Collection

Collecting an emotion dataset is essential for training the system. Data collected for emotion can be divided into three types: acted, spontaneous, and elicited. In acted data, subjects are recorded while performing certain emotions. There are several databases of acted audio and/or video recording available for research studies as reviewed in the Zeng et al. [2009] study. Most of these acted emotions databases are based on the six basic emotions (happiness, sadness, surprise, fear, anger and disgust), which are recognised universally [Ekman, 1999]. Spontaneous emotion datasets are recorded while occurring in real-life settings such as interviews or interactions between humans or between humans and machines. Elicited responses, which aim at inducing an affective reaction, are recorded while, for example, watch-

ing movie clips and pictures [Gunes and Pantic, 2010]. Obviously, differences between spontaneous and deliberately acted emotions exist. For example, differences in appearance and the timing of the emotion. Moreover, acted emotions are inferior and less symmetric compared to spontaneous emotions [Kanade et al., 2000]. Smiles in particular were studied to differentiate between spontaneous and acted smiles. Felt smiles were found to have different movement of muscles, and by observing these differences a real smile could be verified [Frank and Ekman, 1996].

Nevertheless, as far as the affective sensing techniques are concerned, emotion recognition introduces several challenges:

Recording environment: The recording environment and its specification could be a challenge. Recording in a lab situation is different than recording in a real-life situation. The resolution of the voice and/or video recording, as well as the type and quality of the sensors used could influence designing the affective recognition system. Special sensor devises could be used to measure emotions, such as electrode sensors, pressure sensing devices, 3D cameras, Infra Red motion detection, etc. Some of these sensors are physically attached to the subject (e.g. EEG), which could obstruct their movements, while other sensors are not attached to the subject (e.g. camera). However, enhance the utility of the system, and to develop a generalised emotion recognition system, it is preferable to use sensing devises that are available to all users (e.g. camera, microphone) and do not restrict users' movement.

Individual Differences: such as face shape, skin color, hair, wearing glasses, jewelry or make up, and beards could impact the visual processing and, for example, obscure facial features. Facial and eye features, for example, are different for Asians and Europeans in shape and size, which may affect the robustness of the system [Kanade et al., 2000]. Beside appearance differences, individuals express emotions differently in intensity and frequency [Ekman et al., 1997a; Kanade et al., 2000].

Visual Input: challenges include lighting conditions, non-frontal face orientation, scale of the face (close/far) and out-of-plane head motion. Another difficulty is occlusion of the face by a hand, other objects, or even other faces. Suggestions to overcome challenges due to face orientation and occlusion involve having multiple cameras, or statistically trying to predict the missing parts from whatever image information is available [Pantic and Rothkrantz, 2003].

Audio Input: For example, having noisy background, having overlapping speech of several people, the microphone distance from the speaker, are all variables that should be taken into account when collecting emotion data for developing a system that could recognise emotions from speech.

2.2.2 Data Preprocessing

Once the data has been collected, raw data is then preprocessed either by removing noise, reducing the high dimensionality, and/or segmenting interesting parts for feature extraction. Furthermore, data (signal) preprocessing might be used before, during or after feature extraction, which may include:

Segmentation: The data is divided into meaningful parts for further analysis. In speech processing, segmentation could be done to separate speakers' speech, so as to analyse each speaker individually. Such segmentation could be measured automatically using advanced speaker diarisation techniques [Tranter and Reynolds, 2006]. It could also be done by segmenting sentences or utterances for further analysis. In image processing, it could refer to segmenting desired objects from the background such as the face, the eyes, etc. Such segments emphasis on extracting features from meaningful sections rather than using unrelated ones.

Data cleaning or signal enhancement: Enhancing the signal improves the quality of features extracted from the data by applying noise reduction or outlier removal operations. In speech processing, signal enhancement refers to noise reduction or cancellation by performing filtering techniques, or spectral restoration. In image processing, enhancement could include baseline or background removal, de-noising, smoothing, or sharpening.

Normalisation or standardisation: is the process of centering and/or scaling the data in order to compare them. Normalisation can have different meanings based on the application. In statistics, normalisation means adjusting values of different scales to a unified scale. When modeling inputs with different scales, normalisation is recommended [Jayalakshmi and Santhakumaran, 2011]. Normalisation validates the comparison of features from different datasets, since it eliminates variation effects produced by differences in recording environments. In statistics, several normalisation methods could be used (e.g Z-Score, Min-Max, etc.). In audio processing, normalisation could refer to peak normalisation, where the voice volume is normalised for each sample. That is, the volume would be increased or decreased to match a certain signal peak. Generally, in emotion detection, normalisation of the sound is not used as it potentially would remove any emotional variation.

In image processing, the images are normalised to reduce recording differences such as light conditions and illumination. Image normalisation methods include: normalising intensity values or grayscaling the images [Wagner; Hoch et al., 2005], scaling the detected area to a certain size for comparison, and positioning a certain landmark of a detected area to a standard position. For example, scaling and positioning the face area after detection, would standardise the face size even for different distances.

2.2.3 General Feature Extraction

2.2.3.1 Audio Feature Extraction

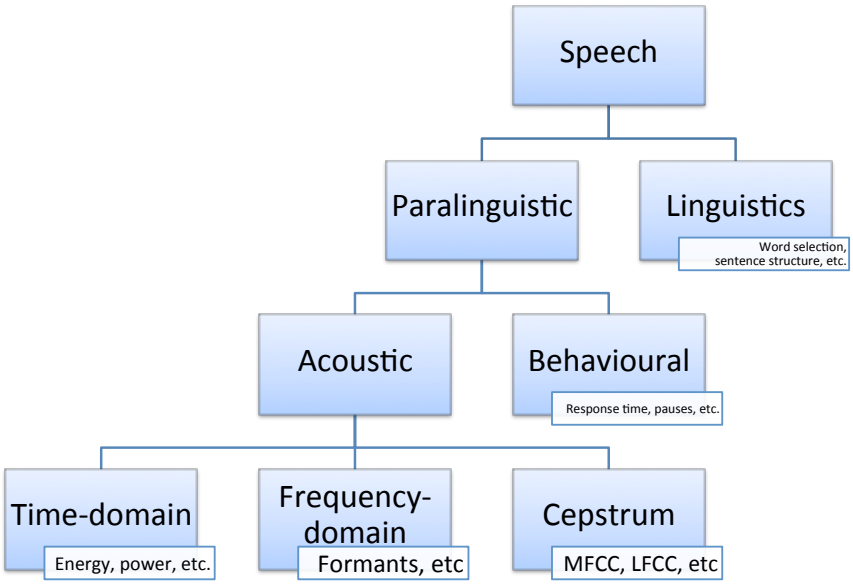


Figure 2.4: General audio feature categories

Speech features can be acquired from the spoken words (linguistic) and from the acoustic cues (paralinguistic) (see Figure 2.4). However, linguistic features, including word choices, sentence structure etc., are not in the scope of this research, especially because I will be analysing depressed speech from different languages (English and German) and focus on generalised objective symptoms that are independent of languages. Paralinguistic features can be divided into speech style features and acoustic features. Speech behaviour or speech style features include, for example, pauses duration and variability, response time, which will be described in detail in Section 4.2.1. Acoustic features extracted from the speech signal include pitch, formants, speaking rate, energy, etc. To extract each type of the acoustic features, the waveform of the speech signal undergoes several pre-processing and signal transformation steps. These are:

Sampling: The conversion of a sound pressure wave as a continuous signal to a sequence of samples as a discrete-time (digital) signal. Typically sound files are sampled at 8kHz or 16kHz for telephone signals, 44.1kHz for CD quality, or 48kHz for professional audio equipment. Such sampling helps in storing the speech in digital systems such as computers. Moreover, knowing the differences in sampling rate between recordings helps in normalising the recordings for comparison.

Pre-emphasing: Increases high frequency components in the signal in order to enhance the overall signal-to-noise ratio. This step is an important pre-processing

step to extract some acoustic features such as formants.

Framing: The decomposition of the speech signal into a series of frames, with preferably overlapping frames, to ensure better temporal continuity. Typically, the frame size ranges are between 10 - 25 ms with 50% or less overlap, which results in 33 to 100 frames per second. Note that some features benefit from a larger frame size such as jitter.

Windowing: Multiplies each frame by a window function in order to improve the frequency-domain representation. Common window functions used in speech processing are Hamming and Hanning [Heinzel et al., 2002]. Moreover, the windowing process reduces the energy leakage generated by the overlapping frames.

Segmenting: The choice of segments depends on the application (e.g. speaker/ speech/ emotion recognition) and the available speech signal types (e.g. spontaneous/ read speech). Nevertheless, speech segmentation could be performed using one or more of the following methods:

Voice Activity Detection (VAD): A technique to detect the presence or absence of human speech. This step avoids extracting acoustic features from silent parts of the speech signal, which could not only affect the accuracy of the system but also unnecessarily increases computational time.

Voiced/Unvoiced: Identifies each frame as voiced (i.e. vibrations in vocal cords) or unvoiced (i.e. no vibrations in vocal cords) based on signal power threshold. Most speech studies investigate voiced signals as they contain rich information about the speaker and speech characteristics. Unvoiced speech also contains information about the characteristics of the speech, such as breathy, whisper, murmur, etc. For example, Clavel et al. [2008] concluded that unvoiced speech enhanced the performance of a fear recognition system.

Utterances vs. Fixed Duration: In general, utterances are identified or separated based on a threshold of the duration of speech and silence in a speech signal. As the name implies, fixed duration segments automatically segment speech signals into a predefined duration of same length (i.e. 500ms). Even though utterances might have better intonation than fixed segments, their segmentation might not be accurate, especially in spontaneous speech, as the utterances may not be complete. Having fixed duration segments could be beneficial for some classifiers in order to get equal vector length for all observation.

Once the speech segments are defined and pre-processd, the windowed frames are used to extract the acoustic features, which can be categoriesed as follows [Wolfel and McDonough, 2009]:

Time-domain features: Time-domain analysis typically yields simple speech parameters efficiently, which are suitable for segmentation purposes. Features extracted or estimated from the time-domain include: energy, zero-crossing, loudness, intensity, duration of speech or silence, and auto-correlation. The auto-correlation feature is a good candidate for estimating F0. Moreover, nonlinear transform of the time-domain is a method to extract Teager energy operator (TEO) features. The TEO is a non-linear speech feature that measures the number of harmonics produced from the non-linear air flow in the vocal tract.

Frequency-domain features: Provide an efficient representation of speech information. The frequency-domain is obtained by converting the time-domain signal using a short-term Fourier transform function. Features extracted based on the frequency-domain include: formants, jitter, shimmer, and HNR. Pitch could be estimated from the frequency-domain, noting that several methods exist for pitch estimation.

Cepstrum features: Cepstral analysis has been widely used for speech analysis. The cepstrum is a Fourier analysis of the logarithmic amplitude spectrum of the signal. Combining the cepstrum with a linear or nonlinear frequency scale is common to extract cepstrum features such as mel-frequency cepstral coefficients (MFCC) and linear-frequency cepstral coefficients (LFCC). The main difference between MFCC and LFCC is that MFCCs are extracted using the mel scale of triangular windows, while LFCCs are extracted using linear scale. Moreover, cepstrum analysis is considered as a reliable way of obtaining an estimate of F0.

The previous categories of acoustic features are extracted at frame level, also referred to as low-level descriptors (LLD). The LLD can be used in the frame-by-frame form for a classifier that deals with low-level features such as a Gaussian Mixture Model (GMM) or Hidden Markov Models (HMM). Otherwise, statistical features such as the average, range, standard deviation, etc. could be calculated over the LLD of a segment to be an input for a classifier such as a Support Vector Machine (SVM).

The low-level features are detailed and rich of information. However, they not only restrict the choice of the classifiers, but also increase the computational time. On the other hand, statistical features summarises the low-level features. Even though the statistical summary might introduce a lose of information, they overcome the low-level shortcomings. In my work, both LDD and statistical speech features are analysed and their performance in depression classification are compared in order to identify which type of features are best to detect depression.

2.2.3.2 Video Feature Extraction

In order to extract features from the video, the Region of Interest (ROI) has to be identified (e.g. face, head, eyes, lips, etc.), located, and tracked (see Figure 2.5).

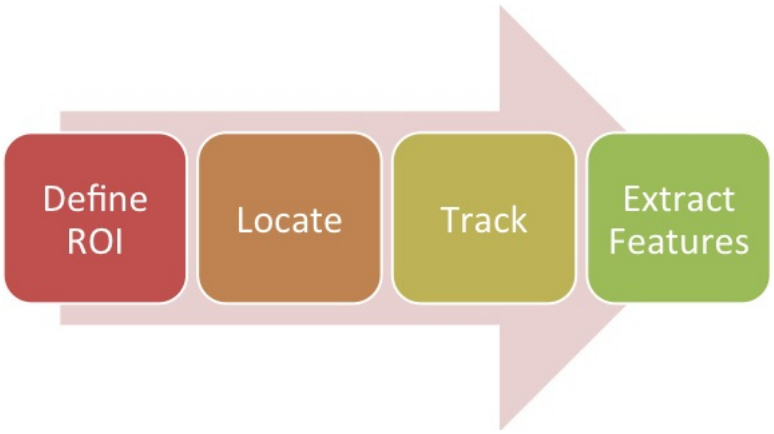


Figure 2.5: General video feature extraction method

Different algorithms could be used for ROI detection, depending how video frames are processed; that is, as a 2D surface or as a 3D volume, as a static single image or as a video sequence. ROI detection techniques can be divided into several approaches [Yang et al., 2002]:

Rule-based: (also referred to as knowledge-based) This approach translates human knowledge about ROI features to rules to be followed by the system. For instance, a face has two roughly symmetric eyes, a nose, and a mouth. Moreover, the relationship between features and the relative distances and position are used as rules to accurately locate the ROI. A disadvantage of this method is that the rules might be too specific or too general, the system may fail to detect ROI or may give false positive detections.

Feature-based: This approach aims for finding structural features, followed by inferring the presence of the ROI. Feature-based approaches detect and localise image features related to the position of the ROI. They rely on the extraction of local features of the region and on fitting the image features to the model. A disadvantage of this approach is that the image could be corrupted due to illumination, noise and occlusion.

Model-based: Model-based approaches do not explicitly detect features but rather find the best fitting model that is consistent with the image [Li et al., 2005].

Motion-based: Motion-based approaches use motion information without any physical ROI information [Gunes and Piccardi, 2009].

Template matching: A pattern in an image patch is used to describe the ROI or the ROI features, then the correlation between the image and the pattern are computed for detection. In other words, this technique searches for areas of an image that match or similar to a template image.

Appearance-based: Unlike the template method, rather than relying on a single instance of the ROI region, the ROI model can be constructed from a large set of training examples with varying pose and light conditions [Hansen and Pece, 2005]. Appearance models detect and track the ROI based on the photometry of the ROI region.

Deformable templates: rely on a generic template, which is matched to the image. It constructs a model in which the ROI is located through energy minimisation, where the model has to be robust to variations of the template and the actual image [Hansen and Pece, 2005], hence the name deformable.

Hybrid: An advanced technique implemented using a combination of the previous methods. For example, using a combination of feature-based and model-based approaches could achieve a good trade off between run-time performance and accuracy [Li et al., 2005]. For instance, the Active Appearance Model (AAM) combines a feature-based method using the ROI texture, a template matching method using the Active Shape Model (ASM), and an appearance-based method using learning techniques to create a face model.

ROI tracking methods can be performed in two ways: either as "tracking by detection", which continuously detect the ROI in the video in every frame as it is an individual image, or by detecting the ROI in the first frame, then tracking it using temporal information. In general, the second type of ROI tracking can be divided into three categories [Zhao et al., 2003]: motion tracking, which tracks the motion of a rigid object; feature tracking, which tracks certain landmark points in the ROI; or complete tracking, which tracks both motion and features. For instance, the AAM mentioned earlier is used to detect and track landmark points in 2D image processing.

Regardless of the method used for detecting and tracking the ROI in a video, the goal is to extract features or descriptors from the ROI for further analysis. The type of extracted features from the ROI depends on the application. To extract features, several studies used the relative distance and movement of the AAM points (e.g. [Abboud et al., 2004]). Optical Flow techniques, which detect relative motion of an object, have also been used for extracting features (e.g. [Yacoob and Davis, 1996]). Another method of extracting features that has been used lately is Space Time Interest Points (STIP). STIP focuses on specific points that are relative to both space and time in an image sequence using Harris corner detection [Laptev, 2005]. Local Binary Pattern (LBP), which is a texture analysis method [Ojala et al., 1996], was extended to videos to extract object features in sequence of frames. This extension introduces the use of LBP features to capture the muscle movements in the face, for example, to recognise certain emotions. Gabor wavelet based features have also been widely used to extract object features.

However, these methods are general for feature extraction, which are not specific for emotions. Several feature types have been used in the emotion recognition literature, where the ROIs mostly are the face or the body (as discussed in the following

section). In my work, the ROIs are the eyes and face in order to extract eye activity and head pose, respectively. The ROIs in my work are detected and tracked using AAM, which is a hybrid method of detecting ROI and tracks the ROI using temporal information, as explained above.

The following section elaborates on specific methods and features that have been used for general emotion recognition.

2.2.4 Features for Emotion Recognition

Vocal expression: Detecting emotions from speech has been widely investigated. Given that not all speech features are related to emotions, the most relevant speech features for emotion recognition should be investigated and analysed. In general, the more relevant features for emotion recognition seem to be duration, MFCC, energy and pitch variation [Batliner and Huber, 2007]. Regarding acoustic features, in a study by Schuller et al. [2007], it was found that duration and energy are the most relevant features to detect emotions. In addition, a study to find trouble in communication in human-computer automatic call center context, found that among duration, energy, and F0 features, the feature that contributes the most in emotion classification was duration [Batliner et al., 2003]. However, Schuller et al. [2007] concluded that using all acoustic features together could obtain better results than a single group. Another study concluded that combining MFCC and Mel-Band-Energy features leads to better emotion recognition, since MFCC features are robust to emotion, while Mel-Band-Energy features were better in recognising fast, angry, slow, and clear speech [Kammoun and Ellouze, 2006].

Facial expression: Most automated facial expression recognition systems use measurements from muscle movement, which was identified by Ekman et al. [1997b]. In their Facial Action Coding System (FACS), sets of Action Units (AUs) identify independent motion of the face, from which an expression could be recognised. By observing which AUs have been accrued, a specific expression is classified and linked to certain emotion. However, manually coding FACS is time consuming, which is considered as an obstacle to the study of emotion. Several computer vision techniques have been proposed to overcome this obstacle, such as optical flow over face region, Principle Component Analysis over full-face to extract Eigenfaces, spatio-temporal properties of facial features, etc. [Fasel and Luettin, 2003; Pantic and Rothkrantz, 2000]. Lately, STIP has been used to extract local facial and body features (e.g. [Song et al., 2013]). LBP-based features have been used for detecting emotion expressions mainly from the face. For example Jiang et al. [2011] used LBP-based features to automatically recognise specific AUs. Gabor wavelet based features extracted mainly from the face were used for emotion detection (e.g. Zhang et al. [1998]).

Body posture: Unlike facial expressions, body posture has no agreed coding system for recognising emotions, where several studies proposed such system (e.g.

[Gunes and Piccardi, 2005]). Studies on body posture used either special sensors, such as a motion capture system, which captures 3D gestures [Kleinsmith et al., 2011; De Silva and Bianchi-Berthouze, 2004], or the Microsoft Kinect, which captures full-body 3D motion [Scherer et al., 2013b; Stratou et al., 2013], or just a standard digital camera [Castellano et al., 2007; Shan et al., 2007; Gunes and Piccardi, 2005]. Regardless of the sensor used, extracted features from the body aim to provide a description about the joints orientation and their distance from each other [Kleinsmith et al., 2011; De Silva and Bianchi-Berthouze, 2004], or by recognising the gesture shape [Gunes and Piccardi, 2009]. For example, Castellano et al. [2007] uses non-propositional movement qualities (e.g. amplitude, speed and fluidity) rather than recognising gesture shapes. On the other hand, Gunes and Piccardi [2009] attempted to recognise body gestures by modeling the body using a combination of silhouette-based and color-based body models, then extracted features such as general changes (e.g. how the centroid, rotation, length, width, and area of the feature increased or decreased), motion and optical flow with respect to a neutral frame. Shan et al. [2007] extracted, without recognising gestures, spatio-temporal features by detecting STIP of body regions in videos.

Head Pose: Analysing head pose for emotion is usually included in the feature extraction methods for body gestures. For example, to detect emotions, Castellano et al. [2007] extracted the head movement velocity with other body movement features, while Kleinsmith et al. [2011] extracted the head position along with joint orientation and distance. Moreover, several studies have found that head pose is a distinguishing feature for most emotions as surveyed by Kleinsmith and Bianchi-Berthouze [2013].

Eye Activity: There has been little research on eye activity that accompanies emotional responses in affective sensing literature. Some studies used special sensors to accurately measure eye gaze and pupil dilation [Lanata et al., 2011; Alghowinem et al., 2014], while others used a simple camera to measure eye activity using computer vision techniques [Li et al., 2005]. Ioannou et al. [2005] included eye features with facial features to recognise emotions using neuro-fuzzy rule based system, obtaining high rates in classification. Moreover, an EEG signal was used for measuring eye activity in response to emotional stimuli (e.g. [Yang et al., 2005]). Generally, the degree of eye opening, gaze direction, fixation duration, pupil dilation, and blinks are the relevant features for emotions.

2.2.5 Feature Selection

Since irrelevant features lead to high data dimensionality and may affect the performance of the system, feature selection techniques can overcome this by selecting relevant features. Generally, feature selection methods can be divided to subset feature selection and feature transformation (see Figure 2.6), which are described as

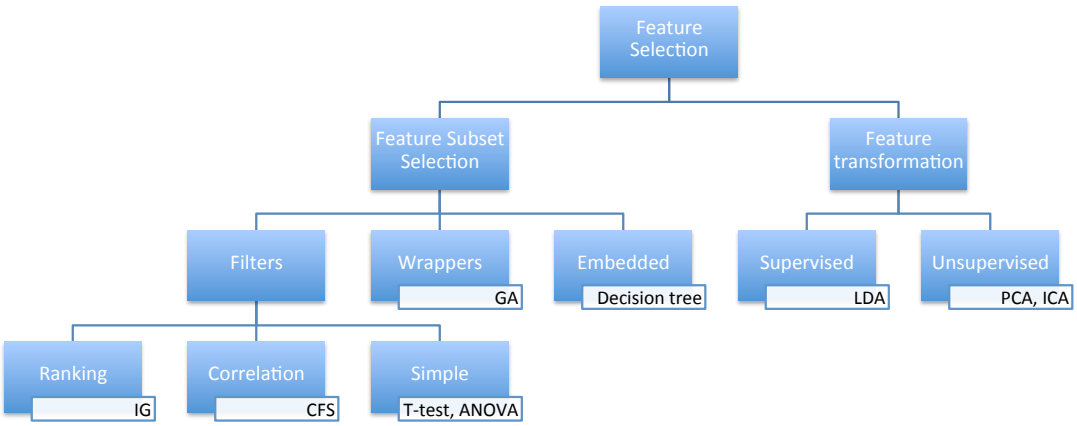


Figure 2.6: General feature selection approaches

follows:

Subset feature selection: is a statistical search technique, finding a relevant subset of features from original features. Statistical methods could be used to compare features and determine the most promising subset. There are a variety of feature selection techniques including, statistical function methods, filter methods, search strategies, learning techniques, etc. In general, feature selection methods can be divided into three categories: filters, wrappers, and embedded [Molina et al., 2002; Guyon and Elisseeff, 2003]. Wrappers and embedded approaches utilise a classifier, such as genetic algorithms (GA) or decision trees, to select the feature subset, which could risk overfitting especially for small datasets. However, filters select a subset of features independently of any induction algorithm based on statistical measures such as ranking, correlation or simple test methods. For example, the Information Gain (IG) filter is a ranking method that estimates the goodness of a single attribute by evaluating its contribution that leads to successful classification [Polzehl et al., 2009]. Another example is correlation-based feature selection (CFS), which finds a subset of attributes that are highly correlated with the class [Hall and Smith, 1998]. Hall and Smith [1998] showed that features that are correlated to the classification problem, but not correlated to each other, increases the accuracy of the classifier output. According to Hall and Smith [1998], CFS uses a search algorithm along with a function to evaluate individual features for predicting the class label along with the level of inter-correlation among them. Nevertheless, both ranking and correlation methods requires a large sample size for a reliable feature selection. Moreover, simple statistical tests such as a t-test and an analysis of variance (ANOVA) test could be used to evaluate the significance of individual features for selection.

Given the small number of observations used in my work, using advanced methods of feature selection risks overfitting to the training set. The overfitting

issue makes it hard to generalise to bigger datasets. Therefore, in my work, a simple t-test is used to filter out the insignificant features as elaborated in Section 3.1.

Feature transformation methods: also known as dimensionality reduction methods, which creates new features from functions of the original features. When the dimensionality of the the original features is very high, dimensionality reduction algorithms are used not only to reduce computational time [Guyon et al., 2006], but also to reduce irrelevant features. For instance, Principle Component Analysis (PCA) reduces dimensionality by projecting the data into a lower dimensional space while retaining as much information as possible using an eigen analysis [Guyon et al., 2006]. Moreover, Linear Discriminant Analysis (LDA), such as Fisher Linear Discriminant Analysis, is used to reduce dimensionality by classifying similar feature data in a supervised method (labels are included in the input). One of the drawbacks of the LDA method is the risk of overfitting to the training set specially when a small data is used. Nevertheless, the output of these dimensionality reduction techniques could be used as the extracted features [Guyon et al., 2006]. Moreover, other general techniques, such as clustering have been used to reduce dimensionality [Guyon and Elisseeff, 2003].

To avoid overfitting when using relatively small datasets, as used in my work, advanced feature transformation methods are not performed. As feature transformation methods, in my work, PCA is used for dimensionality reduction, for its popularity, as well as GMM as a clustering method (see Section 3.1).

2.2.6 Classification

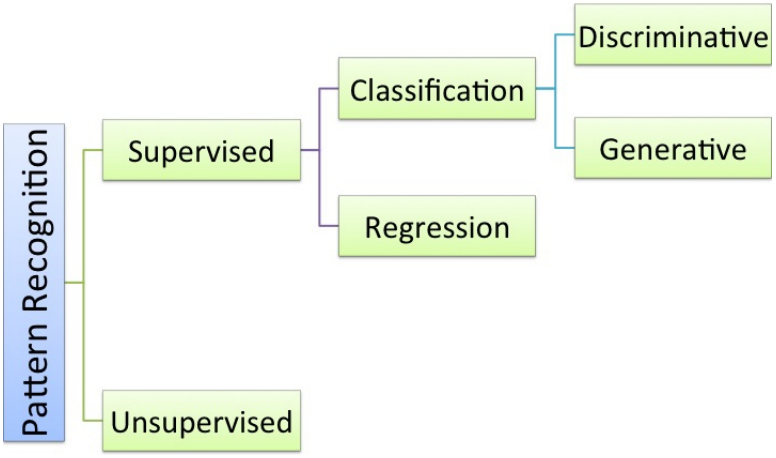


Figure 2.7: General pattern recognition categories

Several classification methods and techniques exist (see Figure 2.7). Based on the system needs, a classifier is selected, which might not be a trivial task. Generally,

the learning methods are divided into supervised and unsupervised learning. In supervised learning, labels are provided with the observations and, therefore, the classification algorithm goal is to generalise to unseen instances. On the other hand, in unsupervised learning, labels are not provided and the classification algorithm's goal is to find a common structure in the observations. With supervised learning, classifiers could predict either discrete classes or a regression value. Most emotion recognition research is based on supervised learning, where the emotions are labeled either by discrete emotion classes (classification), or valence-arousal level (regression) [Zeng et al., 2009; El Ayadi et al., 2011]. Moreover, classifiers could be divided in two categories: generative models and discriminative models. Generative models, such as GMM, learn to cover the subspace that belongs to one class. Discriminative models, such as Artificial Neural Network (ANN) or SVM, learn boundaries between two classes.

The most popular classifiers that have been used in emotion recognition are HMM, ANN, SVM, GMM, and Fuzzy rules [Zeng et al., 2009; El Ayadi et al., 2011]. Moreover, in order to get a balance between classifiers efficiency and accuracy, a hybrid classification has been used in emotion recognition, using a combination of different classification methods [El Ayadi et al., 2011; Datcu and Rothkrantz, 2008; Gunes and Pantic, 2010]. Classifiers used in this work are explained briefly as follows:

Gaussian Mixture Model (GMM): can be regarded as types of unsupervised learning, where the data is clustered and similarities are found. GMM has been widely used in speaker and speech recognition, as well as in recognising emotions [Zeng et al., 2009]. Its advantage is in modeling low-level features directly. GMM can be trained using a continuous Hidden Markov Model (HMM) with a single state that uses weighted mixtures of Gaussian densities [Schuller et al., 2011a].

Support Vector Machine (SVM): is a supervised learning model and considered to be a discriminative classifier. It constructs a hyperplane in a high dimensional space. SVM predicts not only discrete classes in a classification problem, but could also predict continuous values in a regression problem (Support Vector Regression (SVR) [Drucker et al., 1997]). Nowadays, the SVM is considered as a state-of-the-art classifier, since it provides good generalisation properties, even though it might not be the best classifier choice (in terms of generalisability) for every case [Schuller et al., 2011a]. The SVM has been used in both speech and visual emotion classification [Zeng et al., 2009].

Multilayer Perceptron Neural Network (MLP): is a special case of the ANN, which has been used in a wide range of applications, including pattern recognition and emotion recognition. Typically, an MLP consists of an input layer, one or more hidden layers and an output layer, where each layer consists of nodes (perceptrons) that are connected to the nodes of the next layer. MLP networks are usually used for modelling complex relationships between inputs and outputs

or to find patterns in the data. However, the network topology including the number of hidden layers, the number of perceptrons in each layer, the choice of the activation function and the training algorithm are not trivial and complicate the model. Therefore, MLPs are vulnerable to overfitting, typically requiring large amounts of training data [Schuller et al., 2011a].

Hierarchical Fuzzy Signature (HFS): is relatively new, especially to the area of emotion recognition. It overcomes the limitation of fuzzy rules, by handling problems with complex structure and dealing with missing data. Fuzzy signatures can be considered as special, multidimensional fuzzy data. Fuzzy signature compositions data into vectors of fuzzy values, each of which can be a further vector [Tamas and Koczy, 2008]. The fuzzy signature has been successfully applied to a number of applications, such as cooperative robot communication, personnel selection models, and Severe Acute Respiratory Syndrome (SARS) pre-clinical diagnosis system [Mendis and Gedeon, 2008; Ben Mahmoud et al., 2011].

2.2.7 Multimodal Affective Sensing (Fusion Approaches)

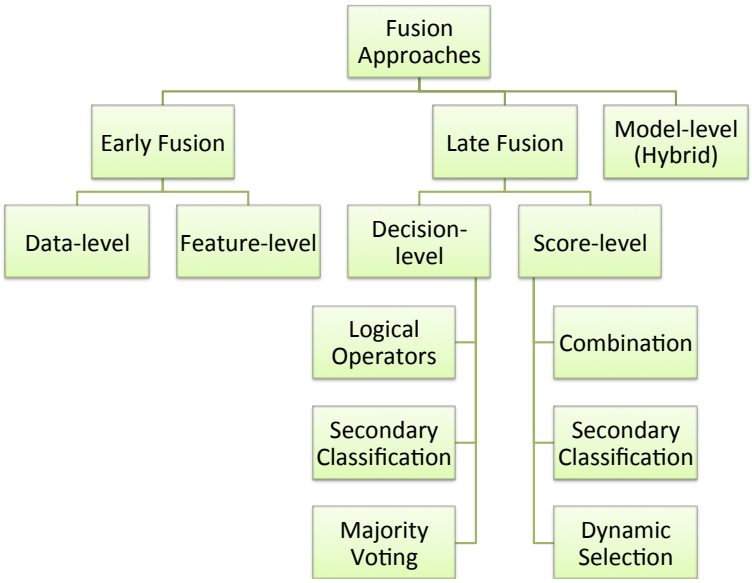


Figure 2.8: General fusion approaches

Logically, a multimodal system that fuses different channels and cues to reach a decision is expected to provide more accurate recognition compared to that obtained using a unimodal system. Several emotion recognition studies investigated fusion approaches to improve the overall recognition results. D'Mello and Kory [2012] analysed some of these studies by comparing their unimodal with multimodal results. Regardless of the considerable variation of these studies, in terms of data, affect, modality, and method, a consistent improvement was found for the multimodal ap-

proach [D'Mello and Kory, 2012]. However, the fusion of different modalities is not a trivial task, as several issues of when and how to fuse the modalities have to be considered [Atrey et al., 2010]. Fusion could be performed as prematching (early) fusion and postmatching (late) fusion, or a combination of both methods as follows (see also Figure 2.8):

Early Fusion: is executed by concatenating the raw data of each sensor (sensor-fusion), or by concatenating the extracted features from the raw data (feature-fusion).

- **Data-level Fusion:** is the integration of raw data from all sensors. This method is difficult to apply to emotion recognition, since the raw data from one sensor has differences in nature, format, sampling time scale and dimensionality from the others (e.g. sounds and images). Besides, the combined data is often huge and computationally expensive and difficult to process to extract features.
- **Feature-level Fusion:** the extracted features of different channels or modalities are combined before classification. This type of fusion is appropriate for synchronised modalities. Studies showed that feature-level fusion gets better results than a decision-level fusion approach [Gunes and Pantic, 2010; Caridakis et al., 2007]. Even though early fusion is expected to contain richer information than late fusion [Rattani et al., 2007], drawbacks of this method are feature vectors from different modalities are not correlated, incompatible, have different time scale and metric levels, and increases feature-vector dimensionality, which might lead to a biased decision towards the larger feature vector. Incompatibility issues have to be fixed before fusing the features using normalisation methods, such as Min-Max and Z-score [Atrey et al., 2010]. Once normalised, features could be simply concatenated, and/or pre-processed for dimensionality reduction by feature selection or feature transformation.

Late Fusion: is executed after the classification of each individual channel, using either the classifiers output scores (score fusion) or labels (decision fusion). Both score and decision fusion could be executed in a simple way (e.g. sum-rule, product-rule, etc., and logical AND, majority voting, etc.), or in a more complex way such as using a secondary classifier [Tulyakov et al., 2008].

- **Decision-level Fusion:** is the fusion of decisions (labels) from each modality's classifier. The fusion is performed either by using operators (e.g. AND, OR), majority voting or a secondary classifier.
- **Score-level Fusion:** is fusing scores from each modality classifier. Fusing scores from different modalities that use the same type of classifier is simple. However, fusing scores from different types of classifiers could be tricky, if the scores are not similar in nature (i.e. distance from hyperplane vs. likelihood ratio). Therefore, further normalisation before the fusion

should be performed [Tulyakov et al., 2008]. Each modality outputs a confidence score, which could be combined using: mathematical operations (e.g. weighted sum, or weighted product), a secondary classifier, or Dynamic Score Selection (DSS) [Rodriguez et al., 2008].

Cues from facial and vocal expressions of emotions are complementary in nature, but also contain some amount of redundancy. Facial and vocal expressions can occur simultaneously or individually. The vast majority of multimodal emotional recognition studies used late-fusion for its simplicity [Zeng et al., 2009].

Model-level (hybrid) Fusion: was used lately, utilising both benefits of early and late fusion [Atrey et al., 2010]. This method of fusion combines both feature-level and decision-level fusion methods to overcome their drawbacks by using the correlation and synchronisation between modalities [Zeng et al., 2009]. A system that uses hybrid strategies including data, feature and decision level fusion is potentially more accurate, flexible, and robust [Dasarathy, 1997].

The concern of overfitting when using a small dataset is an obstacle to investigating advanced fusion methods. However, in my work, early, late and model fusion approaches are investigated. Feature-level as early fusion, several score-level and decision-level as late fusion, as well as hybrid and classifier fusion as modal-level fusion are investigated in Chapter 6 to identify their advantages and drawbacks in detecting depression.

2.2.8 Evaluation and Validation of Affective Systems

The final step in emotion recognition systems is the evaluation, which determines the quality of the system and identifies system weaknesses. That could be done by dividing the dataset into: training, and testing subdivisions. Noting that the training and testing subdivisions should not be overlapping, so as not to contaminate the results [Schuller et al., 2011a]. When only a small dataset is available, cross-validation methods are used to split the training and testing data repeatedly. Cross-validation methods rotate partitions of the dataset to assess how the results will generalise to an independent dataset.

Common types of cross-validation are: k-fold cross-validation, where the original data is randomly partitioned into k equal size subdivisions, and leave-one-out (LOO) cross-validation, which involves using a single observation from the original dataset as the testing data, and the remaining observations as the training data. In emotion recognition in general, using an LOO cross-validation method might present the risk of model contamination, which might occur if the model is trained from observations of the same subject, where it detects the pattern of that subject's observations but not detect the actual problem in hand. To resolve this issue, "leave-cluster-out" cross-validation was introduced by Rice and Silverman [1991], and has been used in emotion recognition literature as "leave-one-subject-out" cross validation, where all

		Actual Label		
		Positive	Negative	Classification Measure
Predicted Label	Positive	TP	FP	$Precision = \frac{TP}{(TP+FP)}$
	Negative	FN	TN	$NPV = \frac{TN}{(FN+TN)}$
	Classification Measure	$Sensitivity = \frac{TP}{(TP+FN)}$	$Specificity = \frac{TN}{(FP+TN)}$	$Accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)}$

Table 2.1: Confusion matrix for a 2-class problem

observations from one subject are left out for each iteration. This cross-validation method has several advantages including preserving within-subject dependency [Xu and Huang, 2012] and reducing model contamination.

Once the model is trained on the training subdivision and then tested on the testing subdivision, the model performance can be measured. Since the analysed data are labeled with the actual classification, the error rate is calculated based on comparing the predicted results given by the classifier(s) with the original labels. A confusion matrix (see Table 2.1) is generated by counting the following for each class in the system:

True Positives (TP): Is a desired result (in the sense of reflecting the true classification), when the predicted result is accepted to be in a particular class, which is the same as the actual label.

True Negatives (TN): Also is a desired result, when the predicted result is rejected to be in a particular class, where the actual label is also not in that class.

False Positives (FP): (also called miss or false reject) Is a misclassified result, when the predicted result is rejected to be in a particular class, where the actual label is in that class.

False Negatives (FN): (also called false alarm or false acceptance) Also is a misclassified result, when the predicted result is accepted to be in a particular class where the actual label is not in that class.

Ideally, a confusion matrix is sufficient to describe the system performance. However, it could be confusing especially when several system configurations are tested. Also, to reduce the confusion, it is preferable to have a certain measure that shows basic system performance. Therefore, based on the confusion matrix, several statistical measures could be calculated, such as accuracy, precision, sensitivity, F1 measure (the harmonic mean of sensitivity and precision), Negative Predictive Value (NPV), etc. [Schuller et al., 2011a]. These measures differ in how much information they reveal about the system performance depending on the desired performance characteristics of the system.

Furthermore, system performance could be graphed for easier visual reference. Common ways to graph system performance use Detecting Error Trade Off (DET) or Receiver Operating Characteristic (ROC). DET has proven to be an easier and more

practical way to measure the error rate than ROC [Martin et al., 1997]. Once the DET or ROC curve is drawn, the equal error rate (EER) can be identified. The EER is the threshold when the false acceptance rate is equal to the false rejection rate, noting that the system with the lowest EER is the most accurate. Moreover, by adjusting the error threshold, the system can be made to meet the desired performance characteristics.

As mentioned earlier, validation by dividing the datasets into several sections is performed to test the system's generalisability to unseen data. However, such division uses the same dataset, which generally has the same recording environment and conditions. Therefore, it could be argued that the system could be dataset-specific and might not have the ability to truly generalise over new recording conditions. Generalising by applying the system on a new data to overcome overfitting to a specific dataset could provide a measure of the feasibility of getting reasonable results on truly unseen data. Moreover, the labels of the new data have to be mapped to the original dataset, which might have more or fewer classes. If that was the case, the misclassifying rate might increase, which also decreases the accuracy of the system [Truong and Leeuwen, 2007]. For example, generalising an emotion recognition system that is trained on a basic six emotions dataset to a valence-arousal emotion dataset is challenging.

Beside the differences in number of classes, other issues of generalising to a new dataset are differences regarding the recording environment, hardware used, spoken language (i.e. from speech modality), the ethnic group (as in different cultures, faces shapes and colour, etc.), etc. In general emotional studies, cross-corpus generalisation is a very young research area. To the best of my knowledge, only few studies have investigated method robustness on different environments [Schuller et al., 2011b; Lefter et al., 2010; Schuller et al., 2010]. Speech in particular is immensely affected by the recording environment, due to varying room acoustics and different types of and distance to the microphones [Schuller et al., 2010]. The video part has also its obstacles regarding recording environment: lighting condition, frame rate, camera's focal point, type and distance, and resolution and dimension size of the video files. In general, due to all mentioned differences, generalising the system to a new dataset gives a lower accuracy result than for the original data (e.g. [Truong and Leeuwen, 2007; Schuller et al., 2010]). Therefore, normalisation methods have to be performed to eliminate recording environment differences [Schuller et al., 2010]. In this work, I attempt to mitigate such differences, as described in Section 3.1.

2.3 Depression Recognition Using Computer Techniques

Despite differences in methods, databases, and classifiers used, a few studies have been investigating the automatic detection of depression lately using computer artificial intelligence techniques using either audio or video channels, as well as multimodal channels.

2.3.1 Depression Datasets

Collecting clinical datasets is time consuming and difficult. It involves ethical and legal processes in order to acquire and use the recordings of subjects. A few datasets have been collected and used for automatic detection of depression as follows (see also Table 2.2 for a comparison):

SDC (suicidal, depressed, and control subjects): is a collection of recorded speech from different datasets. Suicidal speech samples were collected from an existing dataset used by Silverman and Silverman [2002], which contained recorded treatment sessions, phone conversations and suicide notes. For a depressed speech sample, two databases were used. The first dataset is Vanderbilt II, which contained both male and female patients recorded while conducting a therapy session, where patients met with the therapists weekly for up to 25 weeks. The second dataset is from a study comparing the effect of therapy methods on depression collected and used by Hollon et al. [1992]. Depressed patients met DSM-IV criteria for depression and ICD-9-CM (International Classification of Diseases, ninth edition, Clinical Modification). For the control group speech sample, therapists' speech from the Vanderbilt II dataset was used.

PDBH: The psychiatry department of behavioural health at the Medical College of Georgia in the USA collected a database of patients suffering depression (met DSM-IV criteria) and non-depressed control subjects. A total of 15 patients and 18 controls were audio recorded in one session only, while reading a short story.

Pitt: At the University of Pittsburgh (Pitt), a clinically validated depression dataset was collected during treatment sessions of depressed patients. All participants in the dataset were recruited from a clinical trial, where they all met DSM-IV criteria for major depression. A total of 57 depressed patients were evaluated at seven-week intervals using the HRSD clinical interview for depression severity. Interviews were audio-video recorded on up to four occasions and depression severity was evaluated each time by a clinician (see Section 3.2.2 for more details).

BlackDog: The Black Dog Institute, a clinical research facility in Sydney, Australia, collected a clinically validated depression dataset. Audio-video recordings from over 40 depressed subjects with over 40 age-matched controls (age range 21-75yr, both females and males) have been collected. The audio-video experimental paradigm contains several parts, including reading sentences and interviews (see Section 3.2.1 for more details).

Mundt et al. [2007]: collected a dataset for studying depression severity. A total of 35 patients (met DSM-IV criteria and HRSD of > 22) referred by a treating physician to sign a consent and to get access to a touch-tone telephone interface every week for 7 weeks. Each week, speech sample were obtained over

Dataset	Task	# Subjects	Age Group	Depression Scale	Recording	Procedure	Sessions	Language
SDC	depressed	59	adults	DSM-IV & ICD-9-CM	audio-video	therapy session (patient)	several	American English
	suicidal	22			audio-video/ audio only	therapy session, phone conversation, suicidal notes		
	controls	34			audio-video	therapy session (therapist)		
PDBH	depressed	15	adults	DSM-IV	audio only	reading short story	1	American English
	controls	18						
Pitt	depression severity	57	adults	DSM-IV & HRSD	audio-video	HRSD clinical interview	up to 4	American English
BlackDog	depressed	> 40	adults	DSM-IV	audio-video	watching clips, reading speech, structured interview	1	Australian English
	controls	> 40						
Mundt et al. [2007]	depression severity	165	adults	DSM-IV & HRSD	audio only (telephone)	free speech, reading speech	2	American English
WCGS	depressed	16 sever, 52 mild	elderly	CESD	audio only	structured interview for personality	1	American English
	controls	1114						
ORI	depressed	68	adolescents	DSM-IV	audio-video & other sensors	family interaction	1	American English
	controls	71						
ORYGEN	depressed (prediction)	15	adolescents	conventional diagnostic tests	audio-video	family interaction	1	Australian English
	controls	176						
DAIC	depressed, distress, anxiety, total	29% 32% 62% 167	adults	PHQ-9 (self-report)	audio-video	virtual-human interaction	1	American English
AVEC	depression severity	150	adults	BDI (self-report)	audio-video	human-computer interaction (read, free, singing speech)	1	German

Table 2.2: Comparison of datasets used for automatic depression detection

the telephone, which consisted of free speech, reciting alphabets and numbers, reading passage and sustained vowels.

WCGS: The Western Collaborative Group study collected a speech dataset of 1182 elderly men (mean age of 70). Although, the dataset focused on personality types, it also contained data on the psychological assessments of each participant including the depression scale using the Center for Epidemiologic Studies Depression Scale (CESD) [Devins and Orme, 1985]. The data contained 16 severe depressed, 52 mild depressed and 1114 non-depressed subjects. The audio recordings consisted of 15 minutes structured interviews for personality type.

ORI: The Oregon research center in USA collected an adolescents depression audio-video and other sensors corpus. The dataset consisted of recordings of 139 adolescents (12-19 years old) with their parents participating in three types of family interactions (event-planning, problem-solving, family consensus). Of this sample, 68 adolescents met the DSM-IV for major depression, while the remaining 71 participants were healthy non-depressed controls.

ORYGEN: is a youth health research centre in Melbourne, Australia. Before the data collection, a large sample of 2479 potential participants was screened using three different questionnaires and assessments including depression. Only participants diagnosed with no depression were invited for a laboratory recording session that involved interaction with their family. Of this sample, 191 families with their 12-13 year old children participated in the audio-visual recording. Two years after the recording, a follow-up stage was conducted. The second stage involved only a test by psychologists to determine the participants' mental state using conventional diagnostic tests, including screening for depression. At the second stage, it became apparent that 15 participants developed major depression (6 males and 9 females).

DAIC: The virtual human distress assessment interview corpus is an audio-video recordings of participants interacting with a virtual human. Before the recording, participants complete a series of questionnaires that include assessments for depression using PHQ-9. A total of 167 subjects were interviewed, where 29% were diagnosed with depression.

AVEC: is a subset of the German audio-video depressive language corpus (AVDLC) which includes 340 video recordings of 292 subjects, while interacting with a human-computer interface [Valstar et al., 2013]. The participants were recorded 1 to 4 times, with a period of 2 weeks between assessments. The recordings included: sustained vowels, task solving, counting, reading, singing, etc. (see Section 3.2.3 for more details). The AVEC subset contained 150 sessions divided equally for training, development and testing sets.

Table 2.2 summarises and compares the existing datasets used for automatic depression detection. As can be seen, each dataset has a different recording environment and procedure, a different task and depression assessment, and a different

number of subjects and sessions. Some of these datasets were not aimed for depression analysis, as in SDC and WCGS, which implies that the recording techniques might not be sufficient for such analysis. For example, SDC suicidal note recordings were obtained from telephone and tape recordings. Also, the Mundt et al. [2007] dataset recordings were obtained over the telephone, where the quality might not be sufficient for advanced speech analysis. The procedure for the sessions differ between datasets, where some used audio only from read speech to recorded therapy sessions, while others used both audio and video recordings from clinical interviews and designed interactions.

Moreover, depression assessment differs in these datasets from clinical-assessment to self-assessment using a variety of clinical tests. That is, with the exception of DAIC and AVEC, which used depression self-assessment, the other datasets used different clinical-assessment depression tests such as DSM-IV, HRSD, CESD, and other modified versions of depression scales, as in ORI and ORYGEN. In addition, the tasks of each dataset differ, ranging from comparing depressed to control subjects to evaluating depression severity. Comparing depressed with control subjects also was different in the targeted age group, from adults to adolescents, and elderly. Depression severity was investigated from either different subjects or the same subjects regarding treatment progress. Not only the number of participants of those datasets differ, but also the number of recorded sessions for each subject, ranging from one session to several sessions.

It would be ideal to have a large clinical depression dataset that is publicly available in order to standardise the automatic depression detecting methods. Such ideal dataset could also contain several sessions for depressed patients to monitor their treatment progress and at least one session for control subjects for comparison. However, such public depression dataset is difficult to acquire for obvious ethical and legal reasons to protect participants' privacy. Among the previous existing datasets, only the AVEC dataset could be shared under a privacy agreement. However, the depression assessment used in AVEC does not include a clinical evaluation, which might influence the scale of depression and its automatic evaluation. The lack of such standardised dataset introduces challenges such as results replication, and increases difficulties of developing a generalised system that recognise depression symptoms.

In this work I attempt to reduce this gap by replicating and generalising the proposed system to three datasets: BlackDog, Pitt, and AVEC as shown in Chapter 7. In this project, the BlackDog and Pitt datasets were accessed as part of a long-term collaboration with the Black Dog Institute and Pittsburgh University, respectively, while the AVEC dataset was shared under a privacy agreement for participating in the AVEC depression challenge 2014. While using subjects from the outpatients clinic of these datasets may have resulted in a wider variety of depressive disorders, it is assumed for the purpose of this study that any diagnosis of depression is sufficient for the term depression used in this thesis.

The following sections review the studies that have been investigating the automatic detection of depression from verbal, nonverbal and multimodal approaches using the previously mentioned datasets. Moreover, Table 2.3 summarises these stud-

ies.

2.3.2 Detecting Depressed Speech

There have been several studies on detecting depressed speech characteristics and classifications either based on automatic emotion detection studies or based on psychological investigations. While psychological investigations are concerned with the overall patterns of speech using statistical measurements based on functionals of speech prosody features, affective computing classification could use both forms of frame-by-frame low-level features as well as functionals. Therefore, not only previously mentioned vocal features from psychological studies have been investigated for automatic assessment of depression, but also advanced multidimensional vocal features. I will summarise some of them here.

France et al. [2000] used several speech features to classify depressed and suicidal subjects from healthy control subjects. The study used the SDC dataset and separated subjects by gender. They used statistical feature selection with LDA as a classifier to determine which set of features best describes the different classes. The classification results were satisfying with an average accuracy of 75% for both male and female groups. They concluded that the formant and power spectral density measurements were the best discriminators of class membership in both male and female subjects. However, they introduced potential concerns about the differences in recording devices and environment between the three databases used, which might have affected the results.

Moore et al. [2003, 2004, 2008], in a series of studies, investigated several speech features to distinguish clinically diagnosed depressed patients from control subjects. They used the PDBH dataset to extract prosodic, glottal, and vocal tract features. A quadratic discriminant analysis (QDA) classifier along with different combinations of prosodic measures, vocal tract measures, and glottal measures were used, which resulted in on average 87% recognition accuracy. They concluded that the glottal descriptors are vital components to classify affect on speech in general, particularly depression.

Ozdas et al. [2004] have also used a balanced subset of the SDC database, of 10 subjects in each group of depressed, suicidal, and control. They used vocal jitter and glottal flow spectrum speech features for classification. A maximum likelihood classifier was used, which yielded correct classification scores of 83% on average between all three classes. Also, the authors mentioned that using three different datasets with different recording settings could introduce potential noise that could affect the classification results.

Cohn et al. [2009] attempted to detect severity of depression using vocal prosody, in particular pitch and speaker switch duration. Their analysis used a subset of the Pitt dataset containing audio data for 28 participants divided into two categories: 11 of which had a reduction in depression severity by 50% or more in the following sessions and 17 who did not respond to treatment. Using a logistic regression classifier, an accuracy of 79% was achieved.

Cummins et al. [2011] investigated depressed speech from read material. They used a subset of the BlackDog database containing 23 clinically diagnosed depressed patients and 24 healthy control subjects recorded while reading sentences. Several speech features were extracted including prosodic, as well as detailed and broad spectral information. For classification, GMMs with normalisation techniques and several feature combinations were used. A classification accuracy of 80% was achieved using MFCC and formants.

Trevino et al. [2011] used a subset of the Mundt et al. [2007] database to classify depression severity. They used speech interviews of 35 depressed patients who recently started on depression treatment and continued over a 6 weeks assessment period. They segmented the speech into different phonological speech (pauses, vowels, fricatives, nasals, etc.) and studied their correlation with the subjects' depression level. Also, they divided depression scores into 5-classes for classification using Gaussian models. They measured the classification results using the root mean square error (RMSE), which was less than 2 in all classification combinations with different selections of phone lengths and pause lengths. Based on the correlation and classification results, the study concluded that phone-specific characteristics uncover a strong relationship between speech rate and depression severity.

Sanchez et al. [2011] investigated prosodic and spectral features to detect depression in elderly men. They used a balanced subset of the WCGS dataset of 16 depressed and 16 controls. For classification, an SVM is used with a backward elimination as feature selection method to identify best feature groups for the task. They achieved a depression detection accuracy of 81% with prosodic features, namely pitch and energy.

The above mentioned studies investigated detecting depression in adults, while Low et al. [2011] investigated detecting depression in adolescents using the ORI dataset. They extracted several features including: prosodic, cepstral, spectral, glottal, as well as features derived from TEO. GMM and SVM were used for classification, where the results ranged from 67% to 87% based on the features and gender. They concluded that the features that associated with glottal flow formation are important as cues for clinical depression.

Interestingly, Ooi et al. [2012, 2013] investigated potential early predicting of depression in adolescents up to two years in advance. A subset of the ORYGEN database was used in their research. Based on the later diagnosis, the subjects were divided into two groups, adolescents who developed depression and adolescents who did not show any symptoms of depression. Only 15 subjects were assigned to the first group, and to match that sample, only 15 from the second group were selected. A GMM classifier was used with fusing different prosodic, spectral, glottal, as well as TEO features. Their method resulted in a high accuracy level of 73% for predicting future depression.

A recent study by Scherer et al. [2013a] investigated voice quality features for identifying depression and stress disorders from healthy controls. A subset of the DAIC database was used, having 18 moderate to severe depressed subjects, and 18 non-depressed subjects. Using an SVM classifier, an accuracy for detecting depres-

sion of up to 75% was achieved. They concluded a strong characteristic of depressed speech from voice quality features, which is speaker-independent as well as gender-independent.

Williamson et al. [2013] examined vocal biomarkers of depression in order to automatically detect depression severity, using the AVEC dataset. Formant frequencies and Mel-cepstra based features were extracted and processed for correlation structures. They used GMM based regression analysis created from an ensemble of Gaussian classifiers. The results were encouraging, performing down to 7.4 RMSE for speaker-based adaptation approach.

2.3.3 Detecting Nonverbal Depressed Behavior

On the other hand, recognising depression from nonverbal cues has attracted far less attention compared with studies of verbal cues. Nonverbal cues include facial activity and expression, general movement and posture of the body, head pose and movement, as well as gaze and eye blinks.

Cohn et al. [2009] studied automatic detection of depression severity from facial expression. They used a subset of the Pitt dataset of treatment interviews, which were divided into two groups: 66 interview recordings of severe depressed subjects and 41 interview recordings of low depressed subjects from a total of 51 participants (multiple recordings per subjects, see Section 2.3.1). SVM classifiers were used, which yielded 88% accuracy for manual facial action unit features and 79% accuracy for automatic AAM features.

Maddage et al. [2009] classified depression in adolescents with gender dependent and independent models. They used a gender and class balanced subset of the ORI dataset consisting of 4 depressed adolescents and 4 healthy control adolescents. Gabor wavelet features were extracted at 18 facial landmarks from some video frames from the videos and used to create GMM models. They achieved 78.6% average correct classification, with better recognition using gender dependent models.

Ooi et al. [2011] attempt to predict depression in adolescents within 1-2 years in advance from facial image analysis using the ORYGEN dataset. Only 15 adolescents developed depression symptoms after the 2 year period, which have been used in the study with a matching 15 healthy control adolescents. Eigenfaces and Fisherfaces were used as features with nearest neighbour (NN) as a classifier for person dependent and person independent approaches. They achieved up to 61% accuracy using the person dependent approach.

Stratou et al. [2013] investigated nonverbal behaviour to identify psychological disorders such as depression. They used a subset of the DAIC dataset containing 17 depressed subjects as well as 36 non-depressed subjects. Several features were extracted including: basic emotion expressions, facial action units, and head gesturing. A Naive Bayes classifier was used, which achieved 72% F1 measure in classifying depression for the gender independent model.

Joshi et al. [2012, 2013c], in a series of studies, investigated detecting depression from nonverbal cues. Their subset contained interviews of 30 depressed and 30 con-

trol subjects of the BlackDog dataset. Several features were extracted including facial analysis using Local Binary Pattern, upper body analysis using Space Time Interest Points, and a basic head movement analysis. The Bag-of-Words method was used with different codebook sizes and different cluster centres to cluster the raw features. Several classifiers were compared, where SVM performed best achieving 75.3% accuracy using both LBP and STIP. Moreover, facial analysis alone gave up to 72% accuracy, and body analysis resulted in up to 77% accuracy, while head movement was only analysed statistically giving statistically significant differences.

Joshi et al. [2013a] analysed body movements for the task of detecting depression severity. The study used two subsets of Pitt dataset: (1) 12 subjects with two sessions each for both high and low depression scale, and (2) 18 sessions with high depression scale and 18 sessions with low depression scale not necessarily from the same subjects. Holistic body movement and relative body parts movement were fused along with different codebook sizes and different cluster centres to create the features used for the SVM classifier. Classification performance was 87% accuracy for subset 1 and 97% accuracy for subset 2, which was considered a very high performance.

Scherer et al. [2013b] investigated nonverbal behaviours to identify psychological disorders such as depression using the DAIC database. They automatically extracted eye gaze, smile intensity and duration, and manually extracted self-adapters such as the self-touching duration. The features were analysed statistically only, finding significant differences in depressed patients from smile intensity and hand touching duration.

Moreover, to the best of my knowledge; this last study is the only study that has investigated eye behaviour for automatic depression detection. However, other eye activities such gaze direction and duration as well as blink rate and duration were not investigated. Given the relatively small number of and the narrow scope and methodology of prior research on the nonverbal depression cues, I believe that studies on the automatic detection of depression using nonverbal cues are not yet mature and need further investigation. In particular, eye activity and head pose could potentially be a strong indicator for nonverbal behaviour related to depression. In this work, such cues are investigated and the results are presented in Chapter 5.

2.3.4 Multimodal Depression Detection

Most previous studies on the automatic detection of depression have only investigated a single channel, either from video or audio. Multimodal detection of depression is a very young research area. To the best of my knowledge, only a few studies have investigated multiple channels for this task.

In the Cohn et al. [2009] study, mentioned earlier, facial actions and vocal prosody for clinical depression detection were explored. They used different subsets of subjects for the audio and video analysis. Therefore, they did not investigate fusion approaches for the examined channels, but they anticipated an increase of accuracy with fusion techniques.

Cummins et al. [2013] investigated feature fusion of speech and facial analysis

for detecting depression severity using the AVEC dataset. A GMM-UBM system was used for the audio subsystem and STIP in a Bag-of-Words approach for the vision subsystem, which also were fused at future-level. The improvement from the fused system was not statistically significant from the individual subsystems in detecting depression level, performing an RMSE of 10.62 on testing set. Other fusion approaches were not investigated.

Meng et al. [2013] also investigated fusing facial and vocal expression for detecting depression severity using the AVEC dataset. Motion History Histograms (MHH) were extracted from both the audio and video channels to capture temporal changes in facial and vocal expressions. The MHH from the video channel were processed further using an Edge Orientation Histogram (EOH) and Local Binary Pattern to highlight the dynamic feature details. For regression classification to detect the depression scale, Support Vector Regression [Drucker et al., 1997] and Partial Least Squares (PLS) regression [Wold, 2004] are used and compared. Audio and video channels were fused using weighted sum decision fusion. The fusion result improved slightly from individual channels, performing an RMSE of 10.96 on the testing set.

Scherer et al. [2013c] investigated audio-visual indicators for automatic depression detection using a subset of the DAIC dataset. Recorded participants were split into: 14 subjects with moderate to severe depressed, and 25 non-depressed subjects. Audio features were extracted based on voice quality features (e.g. glottal source) and video features were extracted based on emotion and motor variability, then concatenated using feature fusion. Using an SVM classifier, the fused modalities outperformed individual ones significantly, resulting in 90% accuracy (compared to 51% and 64% for acoustic and visual modalities, respectively).

Joshi et al. [2013b] investigated and compared several multimodal fusion techniques for depression analysis. Using a clinical dataset of 60 subjects, the study compared audio-video fusion methods at feature level, score level and decision level. Bag-of-Audio features for the audio channel and Bag-of-Video features for STIP for the vision channel were extracted for analysis. The results showed a statistically significant improvement in the fused systems compared with individual ones, resulting in on average 80% accuracy.

2.3.5 Cross-cultural Depression Detection

Most previous studies on the automatic detection of depression have only investigated on a single language, culture, or database. Using multiple depression datasets for method generalisations is hard to investigate due to differences in recording environments, recording procedures and depression evaluations, not mentioning ethical, clinical, and legal reasons regarding acquiring and sharing such datasets. To the best of my knowledge, only France et al. [2000] and Ozdas et al. [2004] used different database to collect their depression speech sample. In both studies, a preprocessing procedure was performed to compensate for possible differences in recordings. They used Z-score normalisation [Jain et al., 2005] to reduce the effect of recording

differences. Moreover, the acoustic noise, including background noise was removed. Both studies obtained high classification results for control, depressed and suicidal subjects. They also raised concerns of such recording environment differences that might have an effect on the results even with the normalisation methods. However, these studies used each dataset for a separate class (control, depressed and suicidal), which might affect the classification results. That is, the classifier might separate the classes based on their recording environment characteristics, not on the actual class.

As mentioned earlier, three databases will be used in order to investigate not only the feasibility to generalise the methods for different databases, but also investigate objective symptoms of depression across languages and cultures (see Section 3.2).

Table 2.3: Summary of automatic depression recognition studies

Dataset	Classification Task	Study	Procedure	# Subjects	Modality	Classifier	Results
SDC (Audio)	depressed, suicidal vs. controls	France et al. [2000]	clinical interview, suicidal notes	females: 38 patients, 10 controls, males: 21 major depressed, 22 high-risk suicidal, 24 controls	speech (formant and power spectral density)	LDA	average accuracy of 75%
		Ozdas et al. [2004]	clinical interview, suicidal notes	10 depressed, 10 suicidal, 10 controls	speech (vocal jitter and glottal flow spectrum)	maximum likelihood	average accuracy of 83%
PDBH (Audio)	depressed vs. control	Moore et al. [2003, 2004, 2008]	reading story	15 depressed, 18 controls	speech (prosodic, glottal, and vocal tract)	QDA	average accuracy of 87%
Pitt (Audio-Video)	depression severity	Cohn et al. [2009]	clinical interview	audio: 11 low-depressed, 17 high-depressed	speech (vocal prosody)	logistic regression	audio accuracy of 79% (speaker-dependent)
				video: 51 subjects (66 severe depressed sessions, 41 low depressed sessions)	facial expression (AUs, AAM)	SVM	AUs accuracy of 88%, AAM accuracy 79%
		Joshi et al. [2013a]	clinical interview	(1) same 12 subjects, 2 sessions each, of high/low depression, and (2) 18 sessions high depression, 18 sessions low depression	video (holistic body movement and relative body parts movement)	SVM	average accuracy of 87% for subset (1) and 97% subset (2)
BlackDog (Audio-Video)	depressed vs. control	Cummins et al. [2011]	reading sentences	23 depressed, 24 controls	speech (prosodic, detailed and broad spectral)	GMM	average accuracy of 80%
		Joshi et al. [2012, 2013c]	interview	30 depressed, 30 control	video (facial LBP, upper body STIP)	several including SVM	average accuracy of 75%, facial analysis: 72%, and body analysis: 77%
		Joshi et al. [2013b]	interview	30 depressed, 30 control	Audio and Video with several fusion methods (audio: Bag-of-Audio, video: Bag-of-Video of STIP)	SVM	average accuracy of 85%
Mundt et al. [2007] (Audio)	depression severity (5-classes)	Trevino et al. [2011]	free-speech	35 depressed	speech (speech rate of phone-specific)	Gaussian maximum likelihood	RMSE below 2

continue to next page...

Table 2.3: (Cont.) Summary of automatic depression recognition studies

Dataset(s)	Classification Task	Study	Procedure	# Subjects	Modality	Classifier	Results
WCGS (Audio)	elderly men depression vs. controls	Sanchez et al. [2011]	interview	16 severely depressed, 16 non-depressed	speech (prosodic and spectral features)	SVM	average accuracy of 81%
ORI (Audio-Video)	adolescents depression vs. controls	Low et al. [2011]	family interactions	68 major depressed, 71 healthy controls	speech (prosodic, cepstral, spectral, glottal, TEO)	GMM and SVM	accuracy ranged from 67% to 87%
		Maddage et al. [2009]	family interaction	4 depressed, 4 healthy control adolescents	video (Gabor wavelet)	GMM	average correct classification 78.6%
ORYGEN (Audio-Video)	predicting of adolescents depression	Ooi et al. [2013, 2012]	family interaction	15 subjects developed depression, 15 subjects who did not	speech (prosodic, cepstral, spectral, glottal, TEO)	GMM	average accuracy of 73%
		Ooi et al. [2011]	family interaction	15 subjects developed depression, 15 subjects who did not	video (Eigenfaces and fisherfaces)	NN	average correct classification 61%
DAIC (Audio-Video)	depression, stress vs. controls	Scherer et al. [2013a]	virtual-human interaction	18 moderate to severe depressed, 18 non-depressed	speech (voice quality)	SVM	average accuracy of 75%
		Stratou et al. [2013]	virtual-human interaction	17 depressed, 36 non-depressed	video (basic emotions expression, facial action units, and head gesturing)	Naive Bayes	F1 of 72%
		Scherer et al. [2013c]	virtual-human interaction	14 moderate to severe depressed, 25 non-depressed	audio (voice quality) + video (emotion and motor variability) in early fusion	SVM	accuracy of 90%
AVEC (Audio-Video)	depression severity	Williamson et al. [2013]	reading passage	150 recording sessions	speech (Formant frequencies and Mel-cepstra based)	GMM based regression analysis	RMSE of 7.4
		Cummins et al. [2013]	all	150 recording sessions	speech and video feature-fusion (audio: GMM-UBM, video: STIP)	SVR	RMSE of 10.62
		Meng et al. [2013]	all	150 recording sessions	speech and video feature-fusion (audio: MHH, video: MHH+(EOH+LBP))	SVR	RMSE of 10.96

2.4 Summary

For the purpose of this research, depression is defined as any diagnosis of the common types of depression using universal diagnostic tests. Psychological studies investigated depression symptoms (e.g. losing appetite, and lack of sleep), most of which might not be easily detectable by technical sensors. This work focuses on finding verbal and nonverbal behavioural cues that could be analysed by computer technology, which will aid clinicians in diagnosing depression.

A few studies attempted to detect depression or its severity. However, as can be seen from the reviewed studies, each study not only used a different sample size, diagnostic test, and classification task, but also a different signal processing paradigm including feature extraction, feature selection, classification and performance measurements. These differences make it difficult to compare the results and their conclusions. Moreover, these studies investigated some verbal and nonverbal feature sets based on previous psychological findings, while there are still other features to discover, in my opinion. Extracted features were analysed either by using statistical analysis or by using classification methods. I believe that combining classification tasks with a statistical analysis for depression behavioural features will give an insight for not only diagnosing depression, but also for future work on automatic depression detection.

Studies on multimodal aspects of detecting depression have been limited, and do not include many existing fusion techniques. As can be seen from these studies, detecting depression by fusing different modalities increased the accuracy of the classification. Therefore, in this work, fusion approaches will be compared to investigate their contribution to more robust and accurate diagnosis.

Even though using several depression datasets has its legal and technical challenges, it will be useful to investigate the feasibility of a generalised system for this task. Studies using combined depression datasets are scarce, where normalisation techniques have been performed to mitigate the effect of recording environment differences. However, when using different datasets, the classifier might separate the classes based on the recording environment characteristics not the actual class. Therefore, in this work, I attempt to explore this field of study, not only to validate and generalise the investigated system, but also to investigate cross-cultural and aspects of cross-language depression detection. Such a study will require investigating normalisation techniques, as well as extracted feature types.

In the next chapter, a general overview of the steps and methods used in the experiments of this thesis is given (Chapter 3). Further chapters contain experiments of investigating the most accurate configuration for detecting depression from verbal (Chapter 4) and nonverbal (Chapter 5) cues, as well as investigating fusion methods (Chapter 6) and generalisability of the proposed system (Chapter 7).

Chapter 3

Overall System Design and Datasets

Following the generic steps of building a system to detect emotions mentioned in the previous chapter, designing a system to detect depression is investigated. A general overview of the steps and methods used in the experiments of this investigation is given in this chapter. Used methods of each step of the system design are briefly explained in Section 3.1 and will be detailed in further chapters. The general steps are: dataset acquisition and data preprocessing, feature extraction, selection, and normalisation, as well as multimodal fusion. Classification results and statistical analyses used to measure the system performance are also reported briefly in this chapter. Moreover, as obtaining a dataset is a first step for training and testing of any proposed method, Section 3.2 includes details about the main dataset used in this research, as well as details on two other depression datasets used to validate the proposed method and to test its generalisability.

3.1 System Design, Analysis and Evaluation

In the previous chapter, specifically in Section 2.2, the main steps required to build a system able to recognise expressions of emotions were summarised and reviewed. As seen, a diversity of methods exist, which requires selecting methods that are more suitable to the application and the type of data acquired. In this section, a brief overview of the system design and the methods selected to perform each step of the proposed approach are given. A detailed methodology for each modality is given in its designated chapter; that is, Chapter 4 for verbal modalities and Chapter 5 for nonverbal modalities. Figure 3.1 shows the general design and lists the selected methods for the proposed system to detect depression.

Dataset Acquisition

The first step to train and validate a system that recognises emotions in general is to acquire a dataset collected for that purpose, as discussed in Section 2.2.1. Moreover, to build a system that can detect depression cues, a specifically designed data

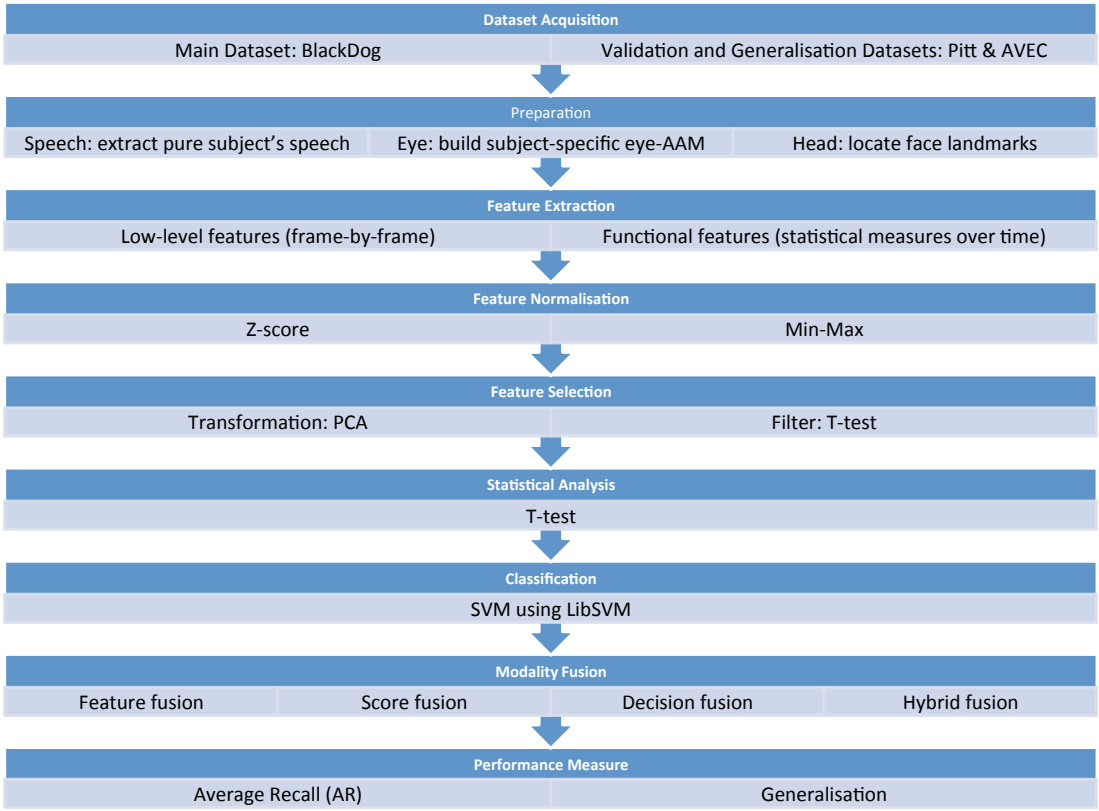


Figure 3.1: Overview of system design and selected method

collection that includes depressed patients has to be obtained. As mentioned in Section 2.3.1, several depression datasets exist and have been used for the automatic detection of depression or its severity. The datasets were originally recorded with different aims in mind, different tests for diagnosing, different experimental paradigm, etc. However, they have similar ethics restrictions and regulations for protecting participants' privacy, which often restricts sharing the datasets. In the field of analysing and detecting depression, using different datasets would enable to test the generalisation capability of the proposed methods and would ensure that these methods are not data-specific. The differences and difficulties in depression data collection could have an effect on building a generalised system that detect depression symptoms. Therefore, in this work, the proposed system to detect depression is not only applied to one dataset, but also validated on other depression datasets.

Even though it would preferable to get access to more datasets and subjects, I was able to obtain ethical permission to use three datasets. These datasets are: Black Dog Institute depression dataset, University of Pittsburgh depression dataset, and Audio/Visual Emotion Challenge depression subset, which were mentioned briefly in Section 2.3.1, and will be elaborated on in detail in Section 3.2. The main dataset

used in this research is the BlackDog dataset since, it has the largest number of subjects and is mainly used for differentiating depressed from control subjects, which reduces variability of depression scales and severity. The BlackDog dataset is used to investigate best approaches to obtain robust and accurate results for detecting depression from verbal and nonverbal cues. To validate and generalise the proposed system, the Pitt and AVEC depression datasets are used, as will be seen in Chapter 7.

Datasets Preparation (Pre-processing)

After acquiring the dataset, several preprocessing steps are required as preparation for feature extraction as briefly explained in Section 2.2.2. The preparation methods used depends on the type of features to be extracted. In this section, the methods used in this research are explained in order to prepare for extracting features from verbal (speech) and nonverbal (eye and head) cues.

In order to extract speech features accurately, the subject's speech should be isolated from other sounds such as noises or other speakers such as the interviewer. This step could be done automatically using an advanced speaker diarisation technique [Tranter and Reynolds, 2006]. However, such techniques are still under development and the error rate would be considered high if an accurate separation is needed. As my work investigates the most distinguishable speech features for depression (see Chapter 4), accurate pure subject speech separation is required to reduce variability of diarisation errors. Therefore, in order to extract pure subject speech accurately a manual approach is used. In the BlackDog dataset, speakers' individual segments are manually separated from overlapped speech and noises, as well as other speech behaviour cues as detailed in Chapter 4. Such intensive separation not only ensures a clean and pure subject speech signal to extract speech features, but also allows extracting speech behaviour features such as interactions between the subject and the examiner including overlapped speech and response time. For the Pitt dataset, the recordings were transcribed manually (details could be found in [Yang et al., 2013]). Using the transcriptions, speakers are separated for pure subjects' speech. On the other hand, as the AVEC dataset does not involve a human examiner, there is only one speaker in each recording. Therefore, no separation is required to extract pure subject speech, and the whole recording is used.

Moreover, once the pure subject speech is extracted, it passes through several speech pre-processing steps such as framing, windowing, and segmenting, as explained in Section 2.2.3.1. For all speech signals, the acoustic features are extracted with frame size set to 25ms at a shift of 10ms and using a Hamming window. For segmentation, a voice activity detector is used to eliminate long pauses within the speech to avoid extracting acoustic features from silent segments of the speech signal. As mentioned in Section 2.2.3.1, identifying the start and the end time of utterances could be used to segment the speech signal. Otherwise, a fixed duration or the entire speech signal could be used. In the next chapter, such segmentation procedures are investigated on the BlackDog dataset in order to determine which segmentation

procedure is favorable in detecting depression.

Unlike the speech signal, the video signal is two dimensional, which contains a lot of information that might not be necessarily useful for each application. Several methods of identifying objects or a region of interest exist, as has been reviewed in Section 2.2.3.2. In this work I am interested in locating and tracking the eyes and the face to be able to extract eye activity and head pose features, respectively.

As the eyes not only occupy a small region in a video frame compared to the rest of the face, but also have very small variations in movements and activity, it is challenging to locate and track them automatically. In order to obtain an accurate location of the eyes as well as accurately track their movements, an Active Appearance Model is trained specifically for the eyes region for each subject using 74 points around the eyes and eyebrows to locate and track the eyes accurately. Specific details on training the eye-AAM are explained in Section 5.1.1.

The same eye-AAM method is used for all three datasets used in this work, the only differences is in the number of annotated images used to train the eye-AAM, which depends on the amount of head movement in the video. For example, as the AVEC dataset used a human-computer interaction paradigm, the subject's head movement is minimal compared to the head movement while being interviewed by a human (BlackDog, Pitt).

Even with its complexity, locating and tracking the face to extract head pose is not as demanding as locating the eyes, because of its larger region and defined structure. To calculate the head pose accurately, several facial landmark points have to be located. With the BlackDog dataset, a user-specific face-AAM is used, which locates 68 points on the face. From these points, the head pose is estimated as detailed in Section 5.2.1. On the other hand, with the Pitt and AVEC datasets a 64-points generic face model using constrained local models (CLM) [Saragih et al., 2009] is used, as detailed in Chapter 7 [1].

Feature Extraction

Once the preparation is accomplished, the next step is to extract the features that are relevant to the application in question. As my work is interested in recognising depression symptoms from verbal and nonverbal cues, I will focus in extracting features that previous studies in the psychology literature proved to be discriminative for depression. Each channel (i.e. audio, and video), and each modality (i.e. speech, eye, and head) has different approaches for extracting relevant features. These approaches will be explained in detail in future chapters, while a brief description is given in this section.

For the speech modality, the extracted features could be divided into two categories: speech behaviour features and vocal prosody features. Speech behaviour

[1] The use of the user-specific face-AAM was aimed for more accurate head pose estimation than generic face models. However, comparing the CLM method with the AAM for head pose estimation on the BlackDog dataset performed similarly. Therefore, CLM was used for the Pitt and AVEC datasets for head pose estimation.

features mostly include duration of speaking, pauses, response time, overlapped speech, etc., which are calculated from the manual labelling of the interviews, as well as other speech rate features, which are estimated using program codes as explained in Chapter 4. On the other hand, vocal prosody features are extracted using open source software over the pure subject speech. The speech signal is framed as 25ms with 10ms overlap, where the features are extracted for each frame, which produces frame-by-frame features, also named low-level descriptors. Unless the system uses specific classifiers that deal with low-level features, functionals or statistical measurements over the LLD could be calculated to be used with different kind of classifiers. In this work, speech behaviour features, low-level features as well as functional features calculated over different segments are investigated, as shown in Chapter 4.

For eye activities as well as head pose and movements modalities, low-level features are extracted (frame-by-frame), as well as several functional features are calculated over specific sections or over the entire session. The frame rate for video is 30 frames per second (fps). Both feature categories for both modalities are investigated and the results are discussed in Chapter 5.

The use of functional features for speech, eye, and head modalities are beneficial for several reasons. First, the session duration is different for each subject. With low-level features, the number of observations depends on the session duration. Unbalanced observations per subject could restrict the choice of classifiers used. Therefore, calculating functionals over fixed or specific sections or over the entire session would ensure balanced observations for each subjects as well as having a wide range of choices for classifiers. Moreover, functional features are robust (with normalisation techniques) for recording environment which will be favorable when using different datasets for generalisation as discussed in Chapter 7.

Worth noting is that the audio and video channels are rich sources of features that potentially have distinguishing characteristics for depression detection. The extracted features in this work are a first attempt to emphasize the importance of these channels, with more features to be investigated in the future.

Feature Normalisation

As mentioned in Section 2.2.2, normalising features of different scales to a unified scale ensures fair comparison. Normalising the data to the same range ensures that each individual feature has an equal contribution to the classification problem, and minimises bias within the classifier for one feature or another. Normalisation is also important when dealing with several datasets, which will minimise the variability of different recording environments.

Moreover, some of the features are also pre-normalised while being extracted using different methods. For example, duration features are pre-normalised over the total duration of the segment or the interview. Low-level features are also pre-normalised. For example, inspired by France et al. [2000] and Ozdas et al. [2004], speech features are pre-normalised using Z-score normalisation for each subject be-

fore extracting the functional feature (see Section 4.3), to ensure a fair comparison for differences from recording environment, subject and gender differences, etc. Eye features are normalised based on other measures to reduce differences due to the distance from the camera and subject differences, as explained in detail in Section 5.1. On the other hand, since the head features are based on the angle pose of the head, pre-normalisation is not required.

The pre-normalised features are further normalised using Min-Max to be scaled similarly to the other extracted features. For all functional features extracted from each modality investigated in this work, normalisation is performed before the classification step in a leave-one-out cross-validation manner. That is, the normalisation parameters are calculated on the training set, and applied to the testing set, where one subject per experiment is left out (leave-one-subject-out).

In this work, performed normalisation methods are:

- Min-Max normalisation, which scales features between 0 and 1, and
- Z-score, which shifts features to a mean of zero and standard deviation of one.

Min-Max and Z-score normalisation have the advantage of preserving exactly all relationships in the data, even though they do not reduce the effects of outliers. The formulae for Min-Max and Z-score normalisation are as follows, given that x is an observation of the feature to be normalised:

$$Min - Max = \frac{x - min}{max - min} \tag{3.1}$$

$$Z - score = \frac{x - \mu}{\sigma} \tag{3.2}$$

More details on pre-normalisation and normalisation will be explained in chapters 4 and 5 based on the extracted features and the investigation in question.

Statistical Analysis

To gain a better understanding of depression characteristics, and to reduce the gap between psychology investigations and affective computing studies, not only pattern recognition techniques are analysed but also statistical analyses. Such statistical analyses are correlated to the classification problem, and could support the automatic recognition results. As mentioned earlier, since the classification is done in a binary manner (depressed/non-depressed, or severe/low depression), a simple T-test is sufficient for finding statistically significant differences between the two analysed groups. Moreover, the statistical analyses of the extracted features give an insight into the contribution of each feature, which would help identifying the behavioural patterns of depressed patients.

As it is statistically difficult to analyse low-level features for significant differences, only functional features are analysed statistically. Furthermore, in order to

characterise depression symptoms, the extracted functional features are also analysed for the direction of the effect.

For the statistical analysis, a two-sample two-tailed T-test are used with all subject data. The two-tailed T-tests are obtained assuming unequal variances with significance $p = 0.05$. The state of the T-test is also calculated to identify the direction of effect.

I acknowledge that a correction for multiple statistical testing (e.g. Bonferroni) might be required. However, as the goal of this thesis is not to identify behaviors that are indicative of depression itself, the correction was not performed.

Feature Selection

As described in Section 2.2.5, feature selection methods reduce irrelevant features in order to enhance system performance. The two main categories of feature selection methods are: (1) subset feature selection and (2) feature transformation, where a method of each of these main categories with the extracted functional features is investigated.

For subset feature selection, a simple filtering method is used. In particular, as the proposed approach is to identify depressed subjects from healthy control subjects in a binary manner (i.e. depressed/non-depressed), using a simple T-test threshold to perform feature reduction is sufficient. That is, features that exceed a statistical threshold set in advance by a t-value corresponding to an uncorrected p-value of 0.05 ($p < 0.05$) are selected for the classification problem (see Section 3.1).

The use of the T-test threshold method as a feature selection is investigated in three approaches. For short, features that exceed an uncorrected threshold of $p < 0.05$ using the T-statistic are abbreviated as ETF, and the approaches are as follows:

All ETF: Using the entire dataset, the T-test is applied on each extracted feature to test its significance between the two groups. Features that exceed the T-statistic threshold are kept fixed for all classification tasks (see Section 3.1). I acknowledge that this method is based on seeing all observations, which might be considered as a contaminated feature selection method. However, to compare classification tasks, a fixed list of features would ensure a fair comparison and verify the robustness of the statistical analysis conclusions.

Variable ETF: In this approach, the features that exceed the T-statistic with leave-one-out cross-validation are selected. That is, using the training subjects in each turn, the T-test is applied to all extracted functional features, then those that exceed the T-statistic with significance $p < 0.05$ are selected for the training and testing sets. Worth noticing is that these features are variable in each run, hence I name this approach variable ETF. Moreover, this approach selects the ETF in each classification task individually. For example, for gender-dependent classification, the ETF are selected based on leave-one-out cross-validation for male and female groups individually.

Mutual ETF: In this approach, the mutual features that exceed the T-statistic in all leave-one-out cross-validations are selected. That is, using the training subjects in each turn, the T-test is applied to all extracted functional features, then only those that exceed the T-statistic with significance $p < 0.05$ in every turn are selected. Then, these mutual ETF are fixed and used for all leave-one-out cross-validation turns. As with the variable ETF, this approach selects the mutual ETF in each classification task individually. Acknowledging the risk that the feature selection is based on seeing all observations, a fixed number of features in each turn of the leave-one-out cross-validation ensures a fair comparison between turns.

The main difference between all ETF and mutual ETF, is that all ETF are selected based on using all extracted features from the entire interviews from all subjects in one turn of T-test. On the other hand, mutual ETF are selected based on a subset of the interviews (i.e. positive and negative questions) or a subset of subjects (i.e males and females) in a leave-one-out cross-validation process.

As a feature transformation method, Principle Component Analysis is used not only for dimensionality reduction, but also to use the most promising principal components (PCs) that have the largest variance, as a method of feature selection. In this work, PCA is performed in a leave-one-out cross-validation manner. That is, the PCA coefficients are calculated on the training set, and then applied to the testing set. The selection of PCs is performed on two approaches:

98% of PC variances: In this method, 98% of PC variances in each leave-one-out cross-validation turn are kept. With this approach, the number of PCs is variable in each turn.

Fixed number of PCs: In this method, the 98% of PC variances are calculated in each leave-one-out cross-validation turn, and then the minimum number of PCs that have at least 98% in each turn are selected. This method would ensure a fixed feature vector length in each leave-one-out turn, thus ensuring a fair comparison between turns.

Classification

Previously in Section 2.2.6, classifiers were explained with a focus on the classifiers used in this work. As mentioned earlier, different classifiers are suitable for different types of features. Therefore, as low-level and functional features are extracted, and to compare the performance of each of these feature types, different methods of classification are required. Unless otherwise stated, for most of the experiments, Gaussian Mixture Models are used with the low-level features, since they have the ability to deal with low-level data of different length, and Support Vector Machines for functional features, since they require an equal feature vector length for all observations. For all classification problems, the classifications are performed in a binary (i.e. depressed/non-depressed, low/high depressed), subject-independent scenario.

To mitigate the effect of the limited amount of data, a leave-one-subject-out cross-validation is used without any overlap between training and testing subsets.

With low-level features, two classification methods are applied. First, using GMM as a classifier. That is, building two GMM models, one model trained on the depressed subjects, and another model trained on the control subjects. These two models do not include the subject that has been left out in the leave-one-subject-out cross-validation. Then, a GMM model is built using the left out subject data and then tested against the two created models. The test calculates the likelihood ratio of the testing model belonging to one of the two models. The classification decision is made based on the likelihood ratio. That is, the smaller the likelihood, the closer the test model is to the group model.

The second method of classifying low-level features is to use GMM as a clustering method for each subject. Then, the GMM model is used as an observation for the SVM classifier. That is, for each subject's low-level data, a GMM is trained with a certain number of mixtures (empirically chosen), then the created model is used as an observation for that subject for an SVM classifier. To select the best number of mixtures (clusters) for the GMM that outputs the most accurate classification results, we experiment on several numbers of mixtures (4, 8, 12, 16, 32) for speech prosody, eye activity and head pose. The highest average of classification accuracy of the three modalities occurred when using 16 mixtures. Therefore, a size of 16 mixtures has been selected and fixed for all further experiments to ensure a fair comparison.

On the other hand, functional features are extracted for each subject to create an observation. These observations are used with the SVM classifier. Several feature selection methods are applied to the functional features as mentioned in the previous section.

For the GMM implementation, the Hidden Markov Model Toolkit (HTK) is used, using one state of HMM to train a GMM with a 16 mixtures and 10 iterations. The number of mixtures is empirically chosen in each experiment and fixed to ensure consistency in the comparison as explained above.

LibSVM [Chang and Lin, 2001] is used for the SVM implementation. To increase the accuracy of SVMs, the cost and gamma parameters are optimised with a wide range grid search for the best parameters with a radial basis function (RBF) kernel. The range is set to -80 to 80 for cost and -80 to 80 for gama, with several steps (40, 20, 10, 5, 2.5, and 0.5) to refine the optimisation. Each step identifies the peak area within its range. Then, the next step searches the identified area from the previous step for another peak and better optimisation, and so on. The final selected parameters are the ones that generalise to all training observations with the leave-one-out cross-validation. In other words, the common parameters that give the highest average training accuracy of all training sets in the cross-validation models are picked, hence the need for wide range search. In an initial experiment, I compared optimising cost and gamma parameters for each training set of the cross-validation models individually, with optimising them based on the highest average training accuracy of all training sets in the cross-validation models. The latter optimisation performed better on the testing sets than individual optimisation. I believe that the individ-

ual optimisation overfit to the training set, while the overall optimisation is able to generalise to the testing sets. Therefore, throughout this work, the optimisation is performed based on the highest average training accuracy of all training sets in the cross-validation turns.

For all modalities, several classification tasks are performed and compared:

Low-level vs. functionals: The performance of low-level and functional features are compared in the classification task to identify, which is more suitable for detecting depression. As discussed previously, different classifiers are used for low-level and functional features, as well as different feature selection methods.

Gender-dependency: While I acknowledge the reduction of sample size as well as the number of observations in this case, gender-dependent classification is performed to investigate the effect of gender in detecting depression. Male and female groups are separated manually and investigated for classification individually.

Expression-dependency: The classification results of different expressions of emotions (positive and negative) are compared. Investigating positive and negative expressions, two related questions are used from the interview that are assumed to elicit the expressions in question. The questions are: "Can you recall some recent good news you had and how did that make you feel?" and "Can you recall news of bad or negative nature and how did you feel about it?". For simplicity, these two questions will be referred to as "Good News" and "Bad News", respectively. I assume that these questions elicit the emotions, even though the answers are not validated for certain emotions. Despite the reduction of sample size (in terms of interview duration), the expression-dependent investigation revealed behavioural pattern differences between the two groups (depressed/controls).

Worth noting is that an initial experiment of classifying depression from speech prosody has been performed on each of the interview questions (8 questions), where the results were compared to the classification result using all interview questions. The initial experiment showed that the highest classification result originated from the data of the "Good News" question and the lowest classification result from the "Bad News" question, while the other questions performed similarly to when using all interview questions.

Since the used datasets are balanced in the number of subjects in each group, the classification chance level is assumed to be 50%.

Modalities Fusion

As elaborated in Section 2.2.7, several modality fusion approaches exist for both early and late fusion categories. For early fusion, a simple concatenation of extracted features and dimensionality reduction of fused features is investigated. For late fusion,

several methods are investigated for score and decision fusion. Moreover, hybrid fusion is also investigated using one and two levels of decision fusion. In this work, each fusion approach is investigated as described in Chapter 6.

Performance Measures

Section 2.2.8 elaborated on measuring system performance by evaluation and validation measurements.

For evaluating the proposed system performance, average recall (AR) is used. The AR, which also called "balanced accuracy", is more informative than the often reported accuracy, as it considers the correct recognition in both groups (depressed/non-depressed). AR is calculated as the mean of sensitivity and specificity, with the following formula:

$$Sensitivity = \frac{TruePositive}{TruePositive + FalseNegative} \tag{3.3}$$

$$Specificity = \frac{TrueNegative}{FalsePositive + TrueNegative} \tag{3.4}$$

$$AverageRecall(AR) = \frac{Sensitivity + Specificity}{2} \tag{3.5}$$

To validate the proposed approach, other datasets are also used to investigate its generalisation properties. The datasets used in this work are described in the following section.

3.2 Datasets

As mentioned earlier, three datasets are used in this study. The main dataset is the BlackDog dataset, which is used for investigations for the most accurate configuration for depression detection. Another two datasets, namely Pitt and AVEC, are used to validate the findings from BlackDog and to test the generalisation ability of the proposed system. The specific details of each dataset are described in the following sub-sections. Also, for easier reference, Table 3.1 summarises and compares the specification of the selected subset of each dataset used in this research.

3.2.1 BlackDog Dataset

The Black Dog Institute is a clinical research facility in Sydney, Australia, offering specialist expertise in depression disorders. Clinically validated data is being collected in an ongoing study. The goal of the Black Dog Institute data collection is to compare subjects diagnosed with different types of depression disorders with healthy control subjects in response to different stimuli. Only subjects who fit the criteria of healthy controls as well as depressed patients are included (see below).

Dataset	BlackDog	Pitt	AVEC
Language	English (Australian Accent)	English (American Accent)	German
Number of subjects	60 (30-30)	38 (19-19)	32 (16-16)
Males-females	30-30	14-24	9-23
Comparison	Depressed/Control	Severe/Low depression	Severe/Low depression
Procedure	open ended questions interview	HRSD clinical interview	human-computer interaction experiment
Symptoms severity measure	DSM-IV	HRSD	BDI
Mean score (range)	19 (14-26)	Severe depression: 22.4 (17-35)/ Low depression: 2.9 (1-7)	Severe depression: 35.9 (30-45)/ Low depression: 0.6 (0-3)
Equivalent QIDS-SR score	19 (14-26)	Severe depression: 17 (13-26)/ Low depression: 2 (1-5)	Severe depression: 20 (16-22)/ Low depression: 1 (0-2)
Total Duration (minutes)	509	355.9	33.2
Average duration per subject (minutes)	8.4 ($\pm$ 4.4)	9.4 ($\pm$ 4.3)	1.0 ($\pm$ 0.8)
Pure subject speech (minutes)	119.3	92.0	23.9
Hardware	1 camera + 1 microphone	4 cameras + 2 microphones	1 web camera + 1 microphone
Audio sampling rate	44100 Hz	48000 Hz	44100 Hz
Video sampling rate	30 fps	30 fps	30 fps

Table 3.1: Summary of datasets specification used in this research

Participants

For depressed patients, only patients who have been diagnosed with pure depression, i.e. those who have no other mental disorders or medical conditions. On the other hand, control subjects are carefully selected to have no history of mental illness and to match the depressed subjects in age and gender. So far, data from over 40 depressed subjects with over 40 age-matched controls (age range 21-75yr, both females and males) has been collected.

Depression Diagnosis and Criteria

Depressed patients were recruited into the study from the tertiary referral Depression Clinic at the Black Dog Institute. All patients were classified as having a current major depressive episode on the Mini International Neuropsychiatric Interview or MINI Sheehan et al. [1998], with the type of depression (variably melancholic, non-melancholic and bipolar depression) rated independently by clinic psychiatrists. Clinical participants were excluded if they met criteria for current and/or past psychosis (unrelated to mood). Additional exclusion criteria, for all participants, included current and/or past alcohol dependence, neurological disorder or history of significant brain injury, a Wechsler Test of Adult Reading (WTAR) Venegas and Clark [2011] score below 80 and/or electroconvulsive therapy (ECT) in the past six months.

Healthy control participants were recruited through the community. Exclusion criteria for healthy controls included current and/or past depression, (hypo)mania or psychosis as assessed by the MINI. All MINI assessments were conducted by

trained research assistants (RAs) at the Black Dog Institute.

Depression severity was assessed using the Quick Inventory of Depressive Symptomatology Self Report Rush et al. [2003], with all clinical participants meeting at least a moderate level of depression severity on this measure. A QIDS-SR score of 11-15 points refers to a "Moderate" level, 16-20 points to a "Severe" level, and ≥ 21 points to a "Very Severe" level.

Data Collection Paradigm

Participants, both depressed and control, are audio-video recorded in one session only. The audio-video experimental paradigm contains several parts, including:

1. **Passive viewing of brief movie clips (1-2 min):** Aiming to elicit a range of content-congruent affective responses, a list of previously validated film clips is used [Gross and Levenson, 1995]. Specifically, the clips used in our paradigm are: Bill Crosby (funny), The Champ (sad), Cry Freedom (anger), Silence of the Lambs (fear), Capricorn One (surprise) and The Shining (fear). This list of clips was selected to elicit an oscillation from positive to negative affect and back. This enables the analysis of facial responses to a variety of controlled positive and negative audio-visual stimuli.

2. **Rating IAPS pictures:** Pictures from the International Affective Picture System (IAPS) [Lang et al., 2005] are presented to participants. Participants are asked to rate each picture as either positive or negative. This enables correlation of the image presentation, the participants' rating and their facial activity.

3. **Reading short sentences aloud:** The reading task contains 20 sentences with negative and positive meaning. For example, "She gave her daughter a slap", "She gave her daughter a doll". This enables the calibration of facial and vocal responses, including the timing of the affective responses to text and semantics.

4. **Answering open ended questions:** An interview is conducted by asking specific open ended questions where the subjects are asked to describe events in their life that had aroused significant emotions. This item is designed to elicit spontaneous self-directed speech and related facial expressions, as well as overall body language. The interview contains 8 question groups, with each group containing some follow up sub-questions. The content of affective situations including neutral situations such as routine activities, positive social events, such as births and weddings, and negative situations, such as bereavement or financial problems, are explored, with a particular focus on perceived mechanisms leading to depression. Open ended questioning was conducted by one of two trained RAs.

Hardware and Specifications

Video and audio streams were captured in QuickTime Pro (running on a 17" Apple Macbook Pro) using a high-resolution Pike F-100 FireWire camera (Allied Vision

Figure 3.2: A view of the BlackDog recording environment setup

Tech.), and broadcast-quality (Sony) lapel microphone. The camera was positioned on a tripod behind the monitor that presented the stimuli, with the height of the camera adjusted for each participant to ensure optimal recording of facial features. The microphone was attached to the participant's lapel, at mid-chest level. During open ended questioning, the RA sat camera-left, behind the monitor (to the right of the participant). Audio was digitised at 44,100 Hz, and the video frame rate was set at 30 fps (frame per second) at 800×600 pixels.

Selected Subset

In this work, a gender balanced subset of 30 depressed subjects and 30 controls is used for the analysis. Only native Australian English speaking participants are selected to reduce the variability that might occur from different accents and language acquisition. For depressed subjects, the level of depression is a selection criterion, with a mean of 19 points (range 14-26 points) of the diagnosis using DSM-IV. I acknowledge that the amount of data used here is relatively small, but this is a common problem in similar studies. As the Black Dog Institute continues to collect more data, future studies will be able to report on a larger dataset.

In this thesis, unless mentioned otherwise, only the interview part of the paradigm is used, as it contains spontaneous interaction behaviour for both audio and video channels. The total duration of the recorded audio-video interview is over 500min (see Table 3.1). Moreover, the interview part is manually labelled to extract pure subject speech, as well as reciprocal speech to extract speech behaviour (see Section 3.1). The total pure speech duration is 290min (see Table 3.1).

3.2.2 Pitt Depression Dataset

At the psychology department of the University of Pittsburgh (Pitt), within a project concerned about treatment of depression, a clinically validated depression dataset was collected [Yang et al., 2013]. The goal of the Pitt data collection is to monitor depression severity during treatment sessions of depressed patients.

Participants

All participants in the dataset were recruited from a clinical trial, where they were all diagnosed with depression. A total of 57 depressed patients were evaluated at seven-week intervals using a semi-structured clinical interview for depression.

Depression Diagnosis and Criteria

In the Pitt dataset, HRSD is used as a measure for assessing severity of depression. HRSD scores of 15 or higher are generally considered to indicate moderate to severe depression; and scores of 7 or lower to indicate no or very low depression [Fournier et al., 2010].

Data Collection Paradigm

The Pitt data collection included interviews using the HRSD questions, where patients were recorded in up to four sessions. Symptom severity was evaluated on these four occasions at weeks 1, 7, 13, and 21 by clinical interviewers.

Hardware and Specifications

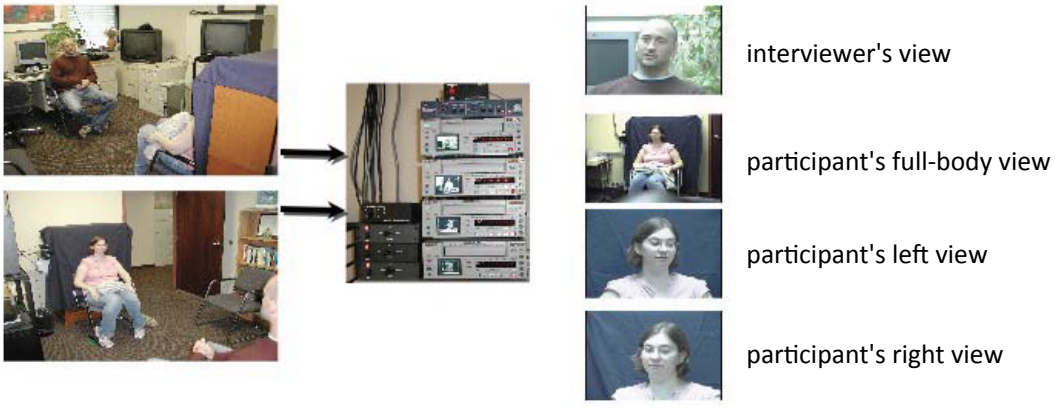


Figure 3.3: A view of the Pitt recording environment setup [Yang et al., 2013]

Interviews were recorded using four hardware synchronised analogue cameras and two unidirectional microphones. Two cameras were positioned approximately

15 degrees to the participant's left and right to record their shoulders and face. A third camera recorded a full-body view while a fourth recorded the interviewer's shoulders and face from approximately 15 degrees to their right. Audio was digitised at $48,000Hz$ and the video frame rate was 30 frames per second at 640×480 pixels.

Selected Subset

In this study, participants from the Pitt dataset are categorised in two categories according to their depression score: low-depression (HRSD score of seven or less) or high-depression (HRSD score of 15 or higher). There are only 19 patients with low depression sessions; therefore, to have a balanced number of participants from both low and high depression groups, another 19 patients with high depression are selected in this study. Since each subject might have up to four sessions, only one session per subject is selected, so that there is no overlap between the subjects from the high and low depression cohort. Video data from the camera to the participant's right were used in the current study.

3.2.3 AVEC German Depression Dataset

The Audio/Visual Emotion Challenge is a subset of the German audio-video depressive language corpus. In its latest version, AVEC included a sub-challenge on automatic estimation of depression level [Valstar et al., 2013]. The data is collected to measure and monitor depression severity.

Participants

The database includes 340 video clips of 292 subjects, with only one person per clip, i.e. some subjects feature in more than one clip. The speakers were recorded between one and four times, with a period of two weeks between the measurements. The mean age of subjects is 31.5 years, with a standard deviation of 12.3 years and a range of 18 to 63 years.

Depression Diagnosis and Criteria

In the AVEC dataset, the depression severity is based on the Beck Depression Index, which is a self-reported 21 multiple choice inventory [Beck et al., 1996]. The BDI scores range from 0 to 63, where 0-13 indicates minimal depression, 14-19 indicates mild depression, 20-28 indicates moderate depression, and 29-63: indicates severe depression. The average BDI-level in the AVEC dataset is 15 points (standard deviations = 12.3).

Data Collection Procedure

The AVEC depression database contains naturalistic video and audio of participants partaking in a human-computer interaction experiment guided by a self-paced Power Point presentation and contains the following tasks:

- sustained vowel phonation, sustained loud vowel phonation, and sustained smiling vowel phonation;
- speaking out loud while solving a task;
- counting from 1 to 10;
- read speech: excerpts of a novel and a fable;
- singing: a German nursery rhyme;
- telling a story from the subject's own past: best present ever and sad event in the childhood;
- telling an imagined story applying the Thematic Apperception Test (TAT) [Murray, 1943].

Hardware and Specifications

The recordings took place in a number of quiet settings and the subjects were recorded by a webcam and a microphone. The audio was recorded using a headset connected to the built-in sound card of a laptop, at a sampling rate of $44,100Hz$, 16 bit. The original video was resampled to a uniform 30 fps at 640×480 pixels.

Selected Subset

In this thesis, a balanced subset of the AVEC database is selected based on the BDI score. The recordings are categorised into binary groups indicating severe-depressed where the BDI score is more than 29, and minimal-depressed where the BDI score is less than 19. Since there are only 16 subjects with a BDI score more than 29, the same number of subjects is selected with ascending low BDI scores from 0 to 4. The spontaneous childhood story telling from the recording tasks is analysed for speech, eye activity, and head pose and movements, in order to mach the spontaneous interview part from the BlackDog and Pitt datasets as closely as possible.

As can be seen, the datasets differ in aim, diagnosis method, and recording procedure/setup. BlackDog aims to compare depressed patients with healthy controls, while both Pitt and AVEC datasets measure and monitor depression severity. The different diagnosis methods made the original depression scores incomparable. However, in this work, each dataset is treated and classified as a binary classification task. That is, with the BlackDog dataset, the system classifies depressed from control subjects, while with Pitt and AVEC, the system classifies high depression from low depression severity.

3.3 Summary

To build a system that could recognise depression, I follow the general steps of emotion recognition systems. In this chapter, I briefly gave an overview of the proposed

system and selected methods for each step. This chapter also included details about the main dataset used in this research, as well as details on two other depression datasets used for the validation and generalisation of the proposed method.

The process of developing an approach that can detect depression from audio and video input contains several steps (see Figure 3.1). One of the first steps is the consideration of which dataset(s) to use. The BlackDog dataset is used as the main dataset to investigate most accurate configuration for detecting depression. The proposed system, derived after extensive experimentation on the BlackDog dataset, is then validated using other datasets, namely Pitt and AVEC, to test its generalisation ability.

Preparing the data for feature extraction is the second step of any system. Each modality has different preparation needs. For example, the speech modality preparation requires extracting pure subject speech from other speakers and noises. On the other hand, eye and head data preparations require locating and tracking the object in question. Although modality preparations could be performed automatically using advanced techniques, manual preparation is used here to obtain accurate results and to reduce error rates that are present even with advanced techniques.

Once the modalities are prepared, features could be extracted. Features are extracted in a low-level format, as well as by calculating statistical functionals over time. The extracted features are then normalised to a unified scale to ensure fair comparison and to eliminate classifier bias for one feature over another. Several features are pre-normalised while being extracted using different methods based on the extracted features. Furthermore, all features are post-normalised using mostly Min-Max (see individual chapters for details).

Features are then filtered to select the most distinguishable feature for detecting depression. Feature selection is performed by a simple filter using features that exceed a T-statistic, as well as by using feature transformation with PCA.

The extracted features are analysed statistically to investigate their significance and direction of effect for the two groups (depressed/non-depressed, or low/high depression). Classification is performed over the low-level and functional features using GMM and SVM, respectively. Before, during or after the classification, fusion techniques could increase the confidence level of the classification results. Therefore, a fusion step is added in the proposed system design, where different methods for each approach of fusion techniques are investigated. Finally, the proposed system performance is measured not only by using "balanced accuracy" (average recall), but also by applying the proposed system on different datasets as mentioned earlier.

For the data acquisition, BlackDog, Pitt, and AVEC depression datasets are used. These selected datasets have many differences in aim, depression diagnosis, depression score, and procedure. However, because acquiring depression datasets has its legal and ethical complexities, I try to overcome these differences. Normalisation techniques and type of extracted features are selected to reduce such differences. For example, as the Pitt and AVEC datasets are aimed at measuring depression severity, the classification task is modified to detect low depression from high depression, in a binary manner.

Future chapters contain experiments of investigating the most accurate configuration for detecting depression. Chapter 4 researches audio channel features and segments that are favourable for depression detection. Chapter 5 focuses on the video channel, specifically eye and head modalities, where the extracted features are analysed and classified. Chapter 6 investigates fusion methods of speech, eye and head modalities. Chapter 7 validates the findings by applying the proposed system to the Pitt and AVEC datasets.

Chapter 4

Depressed Speech Characteristics

Chapter 2 reviewed the verbal and nonverbal symptoms of depression, and also reviewed previous studies investigating automatic depression detection from both audio and video channels. Audio is a rich channel for studying the expression of emotions and mood, not only from spoken content, but also on the speech style and speech characteristics. As elaborated in Section 2.1.2, psychological studies showed that depressed speech has unique characteristics, including speech prosody and speech style features, that differentiate depressed patients from healthy individuals.

In this chapter, using the BlackDog dataset, several speech style and prosody features are extracted and investigated for their effectiveness in detecting depression, in order to address the research question Q1 stated in Chapter 1. This chapter describes the preparation to extract subjects' speech in Section 4.1 and the feature extraction approach, as well as the analysis and classification of the extracted features for both speech style and speech prosody features in Sections 4.2 and 4.3, respectively.

For each speech feature group (style and prosody), several experiments are introduced in this chapter, aiming to explore which speech feature or set of speech features characterise depressed compared to non-depressed speech. First, statistically significant speaking behaviour and functional vocal features that differentiate depressed individuals from controls are explored. Then, the extracted features are used for several classification tasks including: comparing low-level with functional features, investigating gender-dependent classification, and investigating the differences in expressing positive and negative emotions between depressed and control group.

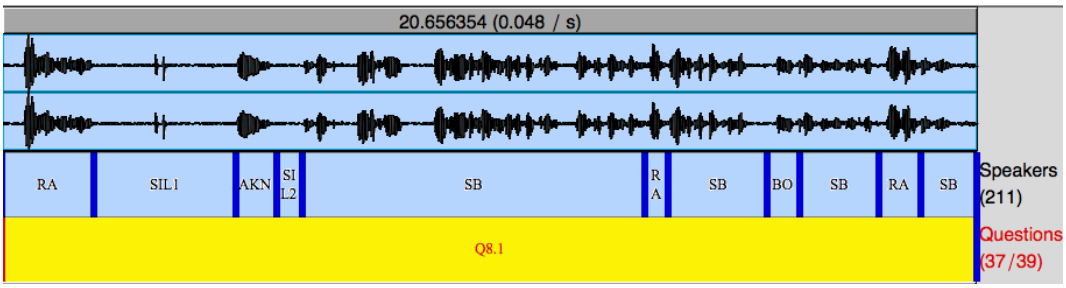
4.1 Speech Signal Pre-processing

Before extracting and analysing speech features, the speech signal has to be cleaned. That is, each subject's voice has to be identified and segmented to remove background noise and remove speech segments from other speakers. As examples of such speech pre-processing, noise removal techniques, which reduce non-speech segments (i.e. background noise) and diarisation techniques, which aim to separate speakers, could be used. Such techniques aim to automate the segmentation process. Some of

the techniques use multiple microphones and then use the differences in signal energy to segment speakers. However, such techniques are not highly accurate and may not be robust to overlapping speech and background noise [Tranter and Reynolds, 2006]. Moreover, the audio recording environment of the BlackDog dataset consists of only one microphone. Therefore, to ensure a highly accurate speaker separation, the annotation of the speech signal was done manually, and then a voice activity detector (VAD) was applied on subjects' speech to measure the sounding and pauses. Further sections detail the process of manual labelling and VAD.

Manual Labelling

Manual labelling can be adjusted to extract features that are applicable to the particular application. More features relevant to speech style and conversational interaction that might not be feasible to extract when using automatic diarisation techniques could be extracted using manual labelling.



RA: Research Assistant speech, SIL1: first silence lag, AKN: Acknowledgment, SIL2: second silence lag, SB: subject speech, BO: overlap speech

Figure 4.1: An example of manual labelling of the BlackDog dataset interview part

For the BlackDog dataset, the interview part, which includes 8 question groups and sub-questions, as described in Section 3.2.1, was manually labelled to separate not only the questions, but also to separate other conversational components within each question such as the speakers (i.e. participant and interviewer). Within each question, reciprocal speech and turns were also labelled (see Figure 4.1), with a focus on the lag between the Research Assistant asking the question and the participant answering it, to measure the response time pattern. In addition, the duration of the overlap between speakers was labelled to measure the involvement style. The duration of subjects' laughs was labelled for further investigation of positive reactions. Also, the duration and the number of the research assistant's interactions to elicit speech from the participants was annotated. Such extensive labelling allows extracting duration features for further statistical analysis of the temporal speech style, as well as accurately extracting subject's speech.

Voice Activity Detector (VAD)

VAD techniques aim to find sounding segments in an audio recording and to reduce silent parts and background noise. This technique is important for extracting speech prosody features, as silent parts would add noise to the extracted features.

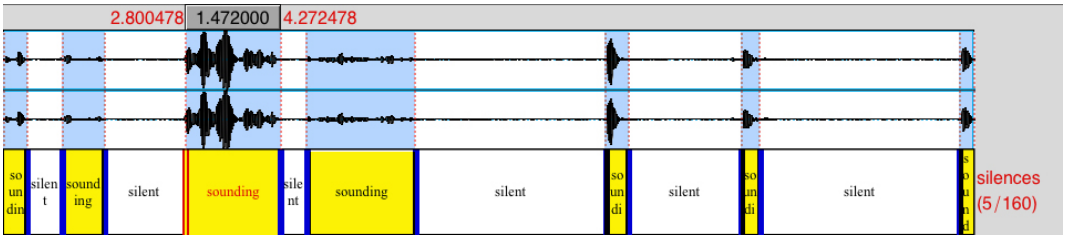


Figure 4.2: An example of applying the voice activity detector on a subject segment. *(Highlighted in yellow are the segments where voice activity is detected.)*

In this work, VAD is applied to the manually labelled subject's speech segments to locate sounding and silent segments. A phonetics software named Praat by Boersma and Weenink [2009] is used to identify voice activity segments by using the intensity of the audio signal. Using an intensity threshold based on the intensity average of each signal, and duration thresholds of silent and sounding, the sounding and silent segments of the audio signal are identified, as shown in the example in Figure 4.2. From the sounding and silent duration, several speech style features were extracted, and from the sounding segments, speech prosody features were extracted, as elaborated in the next sections.

To sum up, in this work, the following segments from the interview part to extract duration features are used:

Subject segments: are all the segments that contain subject's speech (i.e. SB segments from the manual labelling, see Figure 4.1). Note that subject segments include sounding and pauses.

Sounding: are parts of the subject segments where a voice activity is detected, which excludes pauses (see Figure 4.2).

Silence: are the parts of the subject segments where no voice activity is detected. That is, only pauses (see Figure 4.2).

4.2 Speaking Style

Speech is a rich source of emotional cues, not only from the verbal elements, but also from the voice production characteristics such as tune, and from the speaking style such as pauses and speech rate. Moreover, speaking style features could be a good indicator for mood in general and depression in particular. In this section,

several speaking style features are extracted to investigate their ability to distinguish depression disorder.

4.2.1 Extracting Speaking Style Features

In this work, the speaking style feature extraction is divided into two parts: (1) extracting features from the extensive manual labelling, and (2) extracting speech rate features from subjects' segments.

As mentioned earlier, the interview part was manually labelled to separate questions and speakers. Manually labelled speaker turns were used to extract several statistical measurements of the duration for analyses. A total of 63 statistical features are extracted from the manual labelling of the interview, grouped in 7 duration groups:

- subject's speech,
- research assistant (RA) speech,
- time to first response, which is the duration of the silence after asking a question until an acknowledgement indicated by any sounds or words that are not the actual answer for the question (e.g. "ahhh","hmm", "well", etc.),
- total response time, which is the lag between asking the question and the actual answer,
- subject laughing, which indicates a positive affective response in a conversation,
- overlapping laugh, and
- overlapping speech, which measures the involvement style in a conversation.

From each of the above duration feature groups, 9 statistical features are calculated, namely: the average, maximum, minimum, range, variance, standard deviation, total, duration rate (duration of the feature in question $\div$ total duration of the interview), and count (number of occurrences of the feature in question). This resulted in 7×9 features (see top part of Table 4.1).

Speaking rate features were also extracted from subject speech segments by applying VAD using the Praat software, as described earlier in Section 4.1. VAD identifies silent and sounding parts of the subject's speech segments. From the silent and sounding parts, speech, speaking, and pauses duration are extracted as listed in the following feature list. Moreover, to calculate speech rate and articulation rate, the number of syllables has to be calculated, using a Praat script by de Jong and Wempe [2009], which calculates the number of syllables in a sounding segments. A further 19 speech style features are extracted (see bottom part of Table 4.1) as follows:

- Maximum, minimum, range, variance, and standard deviation, for sounding and silent parts (2×5 features),

- For sounding part:
  - number of sounding,
  - total speaking duration (excluding pauses),
  - articulation rate (number of syllable ÷ total speaking duration),
  - average speaking duration (speaking duration ÷ number of syllable),
- For silent part:
  - number of pauses,
  - total silence duration,
  - silence rate (number of pauses ÷ silence duration),
  - average silence duration (silence duration ÷ number of pauses),
- Number of syllables.

It is worth noting that when measuring the speech rate, pauses are included in the duration time of the utterance, while the articulation rate excludes pauses [Goldman-Eisler, 1968].

4.2.2 Statistical Analysis of Speaking Style Features

Speech Style Features		Avg.	Max.	Min.	Range	Var.	Std. dev.	Total Dur.	Rate	Count
From manual labelling	First response	D>C	D>C	D>C	D>C	D>C	D>C	D>C	D>C	-
	Total response	D>C	-	-	-	-	D>C	D>C	-	-
	Overlapped speech	C>D	C>D	-	C>D	-	-	C>D	C>D	C>D
	Laughs	-	C>D	-	-	-	-	-	C>D	-
	Overlapped laughs	C>D	C>D	C>D	C>D	-	C>D	C>D	C>D	C>D
	RA interaction	-	D>C	-	D>C	D>C	D>C	D>C	-	D>C
	Subject's speech	D>C	D>C	-	D>C	D>C	D>C	D>C	-	-
From speech rate	Sounding	-	-	-	-	-	-	D>C	-	D>C
	Silence (pauses)	-	-	-	-	-	-	D>C	-	D>C
	Syllables	NA	NA	NA	NA	NA	NA	NA	NA	D>C

Avg.: Average; Max.: Maximum; Min.: Minimum; Dur.: Duration
Std. dev: Standard Deviation; -: not significant; NA: not applicable (not calculated);
RA: research assistant (interviewer)

Table 4.1: Gender-independent significant T-test results of speaking style features for the interview *(Direction of effect is reported to show which group depressed (D) or controls (C) is higher than the other in the analysed feature, where the $p-value$ shows a significance level < 0.05)*

Speaking style features were analysed for statistically significant differences using two-tailed T-test for two-sample assuming unequal variances and p=0.05. Table 4.1 shows features that exceed the T-statistic ($p < 0.05$) and the direction of effect for the extracted speaking style features listed in the previous section. The direction of effect indicates which group (depressed (D) or control (C)) has a stronger effect. The result analysis is as follows:

- *First response time* – all statistical duration features indicated a longer first response time in depressed subjects. This result is in line with the Zlochower and Cohn [1996] and Ellgring and Scherer [1996] studies, where depressed patients presented longer response times as well. Response latency represents cognitive load that might be caused by psychomotor retardation, which is a main symptom of depression [Parker and Hadzi-Pavlovic, 1996; Parker et al., 1994].

- *Total response time* – the average differences between the two groups were also statistically significant. This shows that the depressed group takes longer time on average to answer the interview questions compared to the control group. The similarity in significance of *first response time* and *total response time* could confirm that this response latency is caused by cognitive load and psychomotor retardation, as well as imply the differences in conversational style and involvement reduction behaviour in depressed subjects.

- *Duration of overlapping speech* – the involvement by depressed subjects was less than by the controls, which is expected for depressed patients [Hale III et al., 1997].

- *Laughter duration* and *overlapped laughter* – the results showed that depression sufferers laughed less. Assuming that a laughter is an expression of general positive affect and social interaction [Gruber et al., 2011], this finding is expected, as negative affect dominates depressed patient emotions [Ekman, 1994; Ekman and Fridlund, 1987].

- Worth noting is that, even though the average *duration of the research assistant speech* was not significantly different between the two groups, the *total duration and the number of research assistant interactions* was higher for depressed patients. The high number of interactions by the research assistant suggests a need for eliciting and encouraging depressed patients to speak more, which may be the reason for having longer speech duration for depressed subjects (*subject's speech*).

- Total *Subject's speech* and *sounding* was longer in depressed subjects. That could be explained by the higher number of research assistant interactions to elicit speech from depressed patients.

- Total and count of *silence duration (pauses)* were significantly longer in depressed patients. As studies showed longer pauses in depressed patients [Ellgring and Scherer, 1996; Reed, 2005; Sobin and Sackeim, 1997], this result is expected.

- *Number of syllables* was higher for depressed group. This finding could be explained by that depressed patients spoke longer in total. However, the *average speaking duration* had no significant differences between the depressed and control groups. This is expected for a spontaneous conversation, as the number of syllables is higher and the duration of a syllable is shorter than for read speech.

Therefore, a conclusion about the cognitive load or about social communication could be derived based on the syllables result.

- *Speech rate* and *articulation rate* was not significantly different between depressed and control groups, which contradicts with the literature where depressed have a slower speech rate [Moore et al., 2004, 2008; Ellgring and Scherer, 1996; Pope et al., 1970]. This contradiction might be caused by the longer duration of depressed subject's speech, which reduced the effect of averaging the speech rate. That is, a longer duration would more likely lead to outliers, that would greatly influence the output of speech or articulation rate.

4.2.3 Classification Using Speaking Style Features

Throughout this work, classification is performed with three tasks in mind, comparing classification results from: (1) low-level (frame-by-frame) with functional features, (2) males with females (gender-dependent), and (3) positive with negative expression (expression-dependent), as described in Section 3.1. Since speech style features are statistical functionals with no low-level features, the classification results of the functional features are shown, which will be a baseline for other classification tasks. These functional features go through different feature selection methods including ETF and PCA. Both ETF and PCA are performed in a leave-one-out cross-validation, to extract variable and fixed features/PCs of each cross-validation turn. Moreover, speech style features were normalised using the Min-Max normalisation to be in a unified scale as described in Section 3.1.

Feature Type	Classifier	Features	Number of Features	Depressed Recall (Sensitivity)	Control Recall (Specificity)	Average Recall
Speech style	SVM	All speech style features	88	76.7	86.7	81.7
		All ETF	45	86.7	90.0	88.3
		Variable ETF	42-49	83.3	83.3	83.3
		Mutual ETF	41	83.3	86.7	85.0
		Fixed PCA variances	27	66.7	90.0	78.3
		98% of PCA variances	27	66.7	90.0	78.3

Table 4.2: Correct classification results (in %) of speech style features

Table 4.2 shows the classification results using different feature selection methods of speech style features. Regardless of the feature selection method, the classification results from speech style features are high, giving 83% AR on average. This high classification result supports previous investigations on speech style of depressed patients, as described in Section 2.1.2.

Even though the differences between the classification results of feature selection methods are not remarkably different, the highest results are obtained by using all ETF and mutual ETF between cross-validation turns. While still being considered a high classification result, PCA performed the lowest, which was also statistically similar to using all speech style features. That implies that PCA was able to represent the feature space with a small number of PCs.

The Table 4.2 also shows the classification recall for both depressed and control groups (sensitivity and specificity, respectively). With all classification and feature comparisons, the recall for controls (specificity) is slightly higher than the recall for depressed (sensitivity), which indicates more depressed patients being misclassified as controls. In this work, there is no threshold set for the classification and an equal error rate is assumed. However, a threshold or a specific error rate could be set based on the application and the desired acceptance and missing rate. Moreover, when using a T-statistic as a filtering mechanism for feature selection, depressed and control recall results are almost balanced compared to when using all features and PCA. This finding supports the usability of the T-test threshold as a feature selection method in a binary classification.

Gender-dependent Classification Using Speech Style Features

To investigate the effect of gender in detecting depression, the classification task is applied on gender-separated cohorts. I acknowledge that this results in a large reduction of sample size as well as in the number of observations. However, such investigation could give an insight not only into gender differences but also into the proposed system's ability to handle a reduced sample size. The number of subjects (observations) and the duration of several segments (sample size) (e.g. sounding, silence) of each gender in each group are shown in Table 4.3.

Duration (min)	Male Subjects		Female Subjects	
	Depressed	Control	Depressed	Control
Subjects	15	15	15	15
Total interview	123.8	112.0	182.5	87.8
Average interview	8.3 (± 3.7)	7.5 (± 2.0)	12.1 (± 6.4)	5.6 (± 1.8)
Total subjects' segments	71.0	62.6	117.2	45.1
Average subjects' segments	4.7 (± 2.8)	4.2 (± 1.5)	7.3 (± 5.2)	3.0 (± 1.3)
Total subjects' sounding	39.3	36.4	69.6	30.4
Average subjects' sounding	2.6 (± 1.6)	2.4 (± 1.0)	4.3 (± 2.1)	2.0 (± 0.9)
Total subjects' silence	31.4	26.1	47.1	14.7
Average subjects' silence	2.1 (± 1.4)	1.7 (± 0.8)	2.9 (± 3.3)	1.0 (± 0.8)

Table 4.3: Total duration of the several segments of the interviews (in minutes) for each gender in each group, as well as average duration and standard deviation (in minutes)

Similar to gender-independent, the gender-dependent classification task is performed using different feature selection methods, as detailed in Section 3.1. Apart from using all extracted speech style features, to ensure a fair comparison with the gender-independent classification, the features that exceeded the t-statistic using all subjects listed in Table 4.1 (all ETF) are kept the same in this gender-dependent classification task. As the T-statistic is used as a filtering method for feature selection, variable ETF and mutual ETF are selected based on the subset used in the classification, which in this case are the male and female groups separately. Moreover, the

PCA, as well as Min-Max normalisation were performed in a leave-one-out cross-validation manner, for each gender-dependent group individually. The results of the gender-dependent classification are shown in Table 4.4.

	Males		Females	
Speech style features	Number of Features	Average Recall	Number of Features	Average Recall
All speech style features	88	66.7	88	80.0
All ETF	45	70.0	45	76.7
Variable ETF	11-19	73.3	44-49	80.0
Mutual ETF	11	80.0	40	86.7
Fixed PCA	19	66.7	19	80.0
98% of PCA variance	19-20	66.7	19	80.0

Table 4.4: Gender-dependent correct classification results (in %) using speech style features

Comparing male and female classification results, regardless of the feature selection method used, female depression recognition results are higher than the ones for males. Only when using all speech style features and PCA, the classification results for females are remarkably higher than for males.

Worth noting is that the number of variable and mutual ETF features that differentiate depressed females from control females is much higher than the number of features for males. This finding might indicate a noticeable difference in the speech style of depressed women, which also might be the reason behind the higher recognition results in females.

The gender differences and the higher recognition results in females are in line with the Nolen-Hoeksema [1987] and Troisi and Moles [1999] studies. Nolen-Hoeksema [1987] concluded that women amplify their mood, while men engage in distracting behaviours that dampen their mood when depressed. Moreover, Troisi and Moles [1999] reported that depressed women showed more socially interactive behaviors than depressed men.

Comparing gender-dependent and gender-independent classification results, the latter are slightly higher. The lower results for gender-dependent classification might be caused by the reduced number of observations. However, even with the reduction of observations in gender-dependent classification, the results were broadly comparable to gender-independent classification, which implies a strong difference between depressed and controls using speech style features.

Positive vs. Negative Expression Classification Using Speech Style Features

As described in Section 3.1, positive and negative expression classification is investigated by using two of the interview questions that were designed to and assumed to elicit the expressions in question. In expression-dependent classification, the same 60 subjects are used as in the general classification. However, the sample size (duration) is shorter. The duration of several segments used to extract speech style features and the number of subjects in each group with each selected question are shown in Table 4.5.

Duration (min)	"Good News" Question		"Bad News" Question	
	Depressed	Control	Depressed	Control
Subjects	30	30	30	30
Total question duration	34.1	21.3	36.1	28.7
Average question duration	1.1 (± 0.7)	0.7 (± 0.4)	1.2 (± 0.9)	1.0 (± 0.6)
Total subjects' segments	18.2	10.4	25.4	16.2
Average subjects' segments	0.6 (± 0.5)	0.3 (± 0.2)	0.8 (± 0.8)	0.5 (± 0.4)
Total subjects' sounding	9.9	6.6	14.9	9.7
Average subjects' sounding	0.3 (± 0.2)	0.2 (± 0.1)	0.5 (± 0.4)	0.3 (± 0.2)
Total subjects' silence	7.1	3.8	9.0	6.5
Average subjects' silence	0.2 (± 0.3)	0.1 (± 0.1)	0.3 (± 0.4)	0.2 (± 0.2)

Table 4.5: Total duration of speech of positive and negative questions from the interview (in minutes) for each group, as well as average duration and standard deviation (in minutes)

To ensure consistency and a fair comparison with previous classification tasks, the same feature selection methods are applied. However, the variable and mutual ETF, PCA and Min-Max normalisation of features are applied on the used subset in a leave-one-out cross-validation method. All ETF are filtered using the entire dataset, which have been listed in Table 4.1. The use of all ETF is to ensure a fair comparison with previous classifications as well as to test their generalisation ability on a smaller subset. The results of the expression-dependent classification are shown in Table 4.6.

Speech style features	"Good News" question		"Bad News" question	
	Number of Features	Average Recall	Number of Features	Average Recall
All speech style features	88	85.0	88	71.7
All ETF	45	76.7	45	70.0
Variable ETF	37-46	76.7	7-15	41.7
Mutual ETF	34	78.3	3	73.3
Fixed PCA variances	27	86.7	24	75.0
%98 of PCA variance	27	86.7	24-26	80.0

Table 4.6: Expression-dependent correct classification results (in %) using speech style features

Comparing the classification results of the "Good News" question with the "Bad News" question (positive vs. negative expressions), the classification results from the positive expression are remarkably higher than the ones for negative expression classification in all feature selection methods. Worth noting is that the segment duration of the "Good News" question is less than the segment duration for the "Bad News" question, and yet the classification results of the positive expression are higher.

As can be seen from Table 4.6, the number of variable and mutual ETF that differentiate depressed from controls while expressing positive emotions is higher than the number of features for the negative expression. This finding implies a bigger difference between the two subjects groups when expressing positive emotion and a smaller difference while expressing negative emotions.

This finding of lower classification results and a lower number of significant fea-

tures, while expressing negative emotions, could indicate that both groups express negative emotions in the same or a similar manner. In contrast, while expressing positive emotions, the classification results and the number of significant features that distinguish depressed from controls are higher. Linking this finding with the previous one, I conclude (1) that positive emotions are expressed less often in depressed subjects, which is in line with Bylsma et al. [2008], (2) that negative emotions dominate in depression sufferers, which is in line with Ekman [1994] and Ekman and Fridlund [1987] and, hence, (3) a negative emotional speech style has less discriminatory power than a positive emotional speech style.

Comparing the results from expression-dependent classification and general classification, general classification results are higher in most cases. The only exception is when using PCA with speech style features for the positive expression, the results are considerably higher than with general classification. Given that the sample size (duration) for each observation has been substantially reduced with expression-dependent classification, the classification results are considered high and comparable to using the entire interview, especially with the positive expression.

At this stage, it is not clear whether the higher classification results for general classification are based on the longer sample size per observation, or based on the combined expressions of emotion data (entire interviews). Even with the large reduction in sample size, the positive expression gave comparable classification results to the general classification case. At the same time, the negative expression had a larger sample size compared to the positive expression, and yet the classification results were remarkably lower. Nevertheless, for future depression dataset collections, a longer duration of positive emotion elicitation interview questions could reveal more distinguishable features and higher classification results than combined emotions or negative ones.

4.3 Vocal Prosody

Speech features can be acquired from both uttered words (linguistic) and acoustic cues (para-linguistic). However, linguistic features including word choices and sentence structure are beyond the scope of this study. I would also like to generalise the findings to other languages in a future chapter, thus, a linguistic investigation would conflict with the generalisation goal of this thesis.

Para-linguistic features include speech style, which have been discussed in the previous section, and include prosody (acoustic) features, which will be investigated in this section.

4.3.1 Extracting Vocal Prosody Features

Prosody features are extracted from sounding segments, such as the energy and frequencies. Prosody features can also be categorised into two branches: low-level descriptors and statistical functionals. Low-level features are extracted frame-by-frame, while functional features are calculated based on the LLD over certain units

(e.g. words, syllables, sentences). In this work, both low-level and functional features are investigated and compared.

To extract the low-level features, sounding segments of a recording have to be identified using voice activity detector. As described in Section 4.1, the intensity function with a calculated threshold is used to identify sounding segments. The sounding segments are then extracted and combined into one audio file for each subject, to prepare for low-level feature extraction. Worth noting is that in a separate work, each sounding segment that lasts over 1.5 seconds was used as an utterance individually for classification (each utterance is an observation) [Alghowinem et al., 2013], where the classification results were not statistically different than when combining those utterances into one file for each subject [Alghowinem et al., 2012]. Therefore, since reliable statistical features need to be computed over longer segments, features from the entire duration of the combined sounding segments are extracted in this work.

To extract low-level features, the publicly available "openSMILE" software by Eyben et al. [2010] is used. Using "openSMILE", the first and second derivatives of each LLD feature are extracted. The low-level features are extracted for each subject with a frame size of 25ms at a shift of 10ms and using a Hamming window, as described in Section 3.1. Knowing that some features benefit from a larger window size such as jitter and shimmer, the window size is fixed for comparison reasons and having a similar number of frames when fusing individual features. In a separate work, using voiced, unvoiced, and mixed speech (both voiced and unvoiced frames) [Alghowinem et al., 2013] were investigated, and I conclude that using mixed speech is more informative, which produces better classification results. Therefore, in this work, mixed speech using both voiced and unvoiced frames is used for all analysis tasks. For each subject, once the low-level features are extracted, they are normalised using Z-score normalisation to reduce recording setting differences between subjects.

To calculate functionals, several statistical measures are applied to the normalised low-level features for each subject regardless of the duration of his/her speech. The statistical functionals include mean, minimum, maximum, and range. Once this process is done, each subject will have one observation of functional features regardless of the duration of their speech. This is beneficial when using an SVM classifier, as the observation length has to be equal, as explained in Section 3.1.

For both low-level and functional features, the most common features in the literature from the fields of psychology and affective computing are extracted as follows:

- The fundamental frequency **(F0)**, which is the lowest frequency of a periodic waveform. Several methods exist to estimate the F0 values; in this work, the Auto-Correlation Function (ACF) is used to estimate F0.

- **Energy** is calculated by both root mean square (RMS) and logarithmic (log).

- **Intensity** and **loudness** are measured as the sum over a short-time frame of the squared signal values. An equal distance from the microphone for all subjects is assumed. Moreover, Z-score normalization for each subject is used to equalize and reduce the variation in the distance from the microphone. That is,

if the distance between each subject and the microphone is different from the other subjects, the Z-score normalization would approximately equalize the difference. Therefore, in doing so, the assumption of equal distance from the microphone for all subjects holds.

- **Jitter** (pitch period deviations) and **Shimmer** (pitch period amplitude deviations) both are calculated using F0, which is estimated using ACF.
- **HNR** is computed from the ACF.
- **Voice probability** is the probability of voicing, computed via ACF.
- **Voice quality** is the output of fundamental frequency quality (the Z-score of ACF).
- The first three **formants** frequencies and bandwidths, which are used to distinguish the vowels in speech and to determine the vowel quality.
- **(MFCCs)**, which are a compact representation of the short-time power spectrum of speech. The first 12 coefficients are extracted by applying overlapping triangular filters.

For all the above features, the first (delta) and the second (delta delta) derivatives are also extracted using "openSMILE". For each subject, the low-level features are normalised using Z-score normalisation to reduce differences between subjects' recordings. Then, the normalised low-level features are used for low-level features classification and also used to calculate the functional features. Moreover, the functional features are further normalised using Min-Max normalisation in a leave-one-out cross-validation manner to be used in the functional features classification task.

4.3.2 Statistical Analysis of Vocal Prosody Features

The use of statistical measures (functional features) makes statistical tests and comparisons feasible between the two groups (depressed and controls). Table 4.7 shows only the features that exceeded the t-statistic.

As shown in Table 4.7, not all extracted features were statistically different between the two groups. That might be due to the subject-specific Z-score normalisation, which reduced the differences between subjects.

F0: The range of the first derivative of F0 was significantly higher in the control group. A lower range of F0 indicates a monotone speech, which is in line with the literature [Nilsonne, 1988; Ellgring and Scherer, 1996; Moore et al., 2004; Mundt et al., 2007; Kuny and Stassen, 1993; Ellgring and Scherer, 1996]. A lower variance of F0 is expected for depressed subjects compared to controls [Moore et al., 2008], however, due to the normalisation process, the differences in the variance of F0 between subjects were cancelled.

Prosody Feature	Derivative	Average	Minimum	Maximum	Range
F0	none	-	-	-	C>D
Voice Quality	none	-	D>C	D>C	D>C
	first	-	C>D	D>C	D>C
	second	-	-	-	D>C
Log energy	none	-	D>C	-	-
	second	-	-	C>D	-
Shimmer	none	-	-	D>C	D>C
	first	-	C>D	D>C	D>C
	second	-	-	D>C	D>C
1st formant frequency	none	-	-	D>C	D>C
	first	-	C>D	D>C	D>C
	second	-	C>D	D>C	D>C
2nd formant frequency	none	-	C>D	-	-
	first	-	C>D	-	-
	second	-	-	D>C	D>C
3rd formant frequency	first	-	C>D	D>C	D>C
	second	-	C>D	D>C	D>C
1st formant bandwidth	second	-	C>D	D>C	D>C
2nd formant bandwidth	none	-	C>D	-	-
	first	-	C>D	D>C	D>C
	second	-	C>D	D>C	D>C
3rd formant bandwidth	second	-	C>D	D>C	D>C
MFCC1	first	-	C>D	D>C	D>C
	second	-	C>D	-	-
MFCC2	none	-	C>D	-	D>C
	first	-	-	-	D>C
MFCC3	none	-	-	D>C	D>C
	first	-	C>D	-	-
MFCC4	none	-	C>D	-	D>C
MFCC6	first	D>C	-	-	-
MFCC8	second	D>C	-	-	-

- :Not Significant

Table 4.7: Gender-independent significant T-test results of speaking prosody features for the interview *(Direction of effect is reported to show which group depressed (D) or controls (C) is higher than the other in the analysed feature)*

Voice quality: Several statistics from voice quality and its first and second derivative were statistically significant to differentiate the two groups. Worth noting is that the larger the variety of fundamental frequency sounding, the richer the voice quality. The minimum, maximum and range of voice quality is higher for depressed than controls. Moreover, the range of the rate of change for voice quality features is higher for depressed than controls. These results indicate a higher stability on voice quality in controls and less control of vocal cords in speech production in depressed patients, where vocal cord dysfunction has been associated with multiple psychological conditions, including major depression [Lacy and McManis, 1994].

Loudness and intensity: Using the BlackDog dataset, these features were not significantly different between the two groups, even though they were expected to be decreased in depressed patients [Scherer, 1987].

Energy: It is expected for depressed patients to have lower energy [Ozdas et al., 2004]. However, most of the statistical measures using the BlackDog dataset

were not significantly different between the two groups, which also might be due to the normalisation process. The minimum of log energy was significantly higher in depressed than in controls. The maximum acceleration of log energy was significantly lower in depressed, which could indicate monotone speech.

Jitter: None of the jitter statistical functionals were significant between the two groups using the BlackDog dataset, even though, according to the literature, jitter is expected to be higher in depressed patients [Scherer, 1987; Nunes et al., 2010].

Shimmer: In line with the literature, the range of shimmer is lower in depressed subjects [Scherer, 1987; Nunes et al., 2010]. Moreover, several statistical measures of the first and second derivatives of shimmer were statistically significant between the two groups, which indicates a strong difference in amplitude variability in depressed patients from control subjects. These differences, as with voice quality, show stability in controls and less control of the vocal cords in depressed subjects.

HNR: A higher HNR is expected for depressed patients [Low et al., 2011], although using the BlackDog dataset, none of the HNR statistical functions were significant between the two groups.

Formants: Most statistics of the first three formants frequencies and bandwidths were significantly different between the two groups. Worth noting is that for the first three formants frequencies, the minimum was higher in depressed than in controls. On the other hand, the first three formants' bandwidths and their first and second derivatives had statistical significance between the two groups. In line with the literature, the results show a decrease of the first three formants' bandwidths [Flint et al., 1993; Moore et al., 2008].

MFCC: Even though MFCC are used for speech and speaker recognition, being relatively robust against noise as well as being dependent on the spoken content, they have been proven to be beneficial for emotion recognition as well. MFCC are used mostly as low-level features. Only few emotion studies used statistical measures from MFCC (e.g. [Kwon et al., 2003; Schuller et al., 2005; Vogt and Andre, 2005; Bitouk et al., 2010]). As can be seen from Table 4.7, several statistical measures derived from MFCC coefficients and their derivatives were statistically different between the two groups.

4.3.3 Classification Using Vocal Prosody Features

Most of the classification tasks in this work investigate the differences between using low-level and functional features, between genders, and between positive and negative expressions. Since several prosody features are extracted, their classification is also investigated individually and when fused. Moreover, several feature selection methods are used with the functional features. The results from these classification tasks and feature selection methods are described in following sections.

Low-level vs. Functional Prosody Features Classification

The extracted prosody features are used for classification individually in both low-level and functional form. The individual feature classification results are shown in Table 4.8.

Feature Type	Low-level		Functional
Classifier	GMM	GMM+SVM	SVM
Log energy	53.3	61.7	60.0
RMS energy	55.0	60.0	61.7
Loudness	46.7	70.0	55.0
Intensity	56.7	61.7	46.7
Shimmer	65.0	65.0	68.3
Jitter	40.0	70.0	00.0
HNR	61.7	61.7	55.0
Voice prob.	36.7	68.3	00.0
Voice quality	56.7	70.0	61.7
F0	46.7	66.7	78.3
Formants	61.7	68.3	71.7
MFCC	66.7	70.0	58.3

Table 4.8: Correct classification results (in %) of individual speech prosody features

Comparing individual prosody feature classification results using low-level features in both the GMM classifier and the hybrid classifier of GMM and SVM, as well as functional features using SVM, in general, using the hybrid GMM-SVM classifier performed higher than other classifiers in most cases. Worth noting is that the number of mixtures of GMM was empirically selected and fixed to all individual features for consistency in comparison, knowing that some features might benefit from having detailed modelling. The only two features that had statistically above chance level classification results ($> 60\%$) consistently with all classifiers are shimmer and formants. Using GMM as a classifier, the highest classification result was obtained by MFCC, followed by shimmer, then HNR and formants. All classification results from the hybrid GMM-SVM classifier were above chance level.

On the other hand, using functional features with the SVM classifier, the highest classification result was obtained by F0, followed by the formants, then shimmer. The classification results from the functional features are consistent with their statistical analysis. That is, the features that had significant differences between the two groups had a high (above chance level) classification result. The only exception to this finding is when using RMS energy features, where there was no statistical significant difference detected, yet the classification results were relatively high. That might be due to a correlation between the statistical measures with the classification problem rather than the significance of individual features **?**.

Table 4.9 shows the classification results of low-level and functionals of all extracted prosody features. For functional features, different feature selection methods are used for comparison to using all functional features. As explained in Section 3.1, ETF and PCA are used as a feature selection method with fixed and variable features/PCs.

For low-level features, GMM is used for its ability to deal with low-level features

Feature Type	Classifier	Features	Number of Features	Depressed Recall (Sensitivity)	Control Recall (Specificity)	AR
Low-level	GMM (16 mixtures)	frame-by-frame	84	66.7	63.6	65.0
	GMM+SVM	GMM model	16 mixtures	90.0	53.3	71.7
Functional	SVM	All prosody features	504	13.3	100.0	56.7
		All ETF	63	76.7	83.3	80.0
		Variable ETF	52-69	56.7	80.0	68.3
		Mutual ETF	45	80.0	86.7	83.3
		Fixed PCA variances	53	13.3	93.3	53.3
		98% of PCA variances	53	13.3	93.3	53.3

Table 4.9: Correct classification results (in %) of speech prosody features

that are imbalanced in length between observations (i.e. duration in the speech case). GMM is used as a classifier and also as a modelling method (clustering) to be used with SVM. Using GMM as classifier performed statistically above chance level with balanced recalls for both groups. Even though the use of the hybrid classifier of GMM and SVM performed higher than using GMM alone, the sensitivity and specificity results of the hybrid classifier are not balanced, where the control recall is at chance level. That might be due to the number of clusters used (16 mixtures) not being suitable for this case. The number of mixtures was empirically selected and fixed for individual features and fused features for comparison. The obstacle of empirically selecting the right number of mixtures reduces the practical use of GMM in this study.

For functional features, as can be seen, using all features and PCA performed very low and almost at chance level. That might be due to the fact that the PCA is a representation of the feature space, where the feature space contains irrelevant features. Using variable ETF, depression recall was almost at chance level, which makes it an unreliable method of feature selection in this particular case. On the other hand, using all ETF listed in Table 4.7 and mutual ETF of all cross-validation turns, performed better than the maximum results of individual features. That shows the contribution of fused features in the classification task.

Comparing low-level with functional features, generally, functional feature classification results were statistically higher than low-level ones particularly with T-test threshold as feature selection. Functional feature classification results using T-test threshold as a feature selection method in both all ETF and mutual ETF performed the highest with balanced recalls in the two groups.

Given that the highest classification results obtained by using fused prosody functional features, and to be consistent with other classification experiments in this work, only functional features will be used in further investigations below.

Gender-dependent Classification Using Prosody Features

To investigate the differences between genders in detecting depression, male and female subjects are separated and then the classification task is applied for each gender individually. The relevant subset duration of sounding segments used to

extract prosody features is shown in Table 4.10 for each gender in each group, which is an extract from Table 4.3.

Duration (min)	Male Subjects		Female Subjects	
	Depressed	Control	Depressed	Control
Subjects	15	15	15	15
Total subjects' sounding	39.3	36.4	69.6	30.4
Average subjects' sounding	2.6 ($\pm$ 1.6)	2.4 ($\pm$ 1.0)	4.3 ($\pm$ 2.1)	2.0 ($\pm$ 0.9)

Table 4.10: Total duration of the subjects' sounding segments of the interviews (in minutes) for each gender in each group, as well as average duration and standard deviation (in minutes) *(extract from Table 4.3)*

The gender-dependent classification results using functional prosody features with different feature selection methods are shown in Table 4.11.

Speech prosody features	Males		Females	
	Number of Features	Average Recall	Number of Features	Average Recall
All speech prosody features	504	60.0	504	76.7
All ETF	63	73.3	63	90.0
Variable ETF	27-50	66.7	97-129	83.3
Mutual ETF	24	90.0	90	90.0
Fixed PCA	26	60.0	26	83.3
98% of PCA variance	26-27	60.0	26	83.3

Table 4.11: Gender-dependent correct classification results (in %) using speech prosody features

Comparing gender classification results, female classification results are statistically higher than in males, with the exception of mutual ETF where the results are equal in both gender groups. The highest result for both male and female groups was obtained when using mutual ETF, giving 90% correct classification, which is considered as a high classification result. Consistent with speech style gender-dependent classification, the female groups have a higher recognition rate than males.

Table 4.11 also shows the number of features selected for classification. Worth noting is that ETF and PCA, both fixed and variable, as well as Min-Max normalisation are performed as leave-one-out cross-validation for each gender group individually. As can be seen when using ETF, the number of selected feature for females is more than treble that in males. That indicates a bigger difference in depressed and control women compared to men.

The higher number of features and the higher classification results for females compared to males shows a noticeable difference between the genders in showing depression symptoms. This finding indicates that depressed women are easier to recognise as they emphasise their mood Nolen-Hoeksema [1987]; Troisi and Moles [1999].

All features and all ETF listed in Table 4.7 are fixed to test their ability to generalise and to be compared with gender-independent classification. Even with the reduction of observations, when using speech prosody features, gender-dependent classification performed statistically higher than gender-independent classification in

all cases of feature selection, with the exception of all and variable ETF for male classification. Even though the subject-specific Z-score normalisation also reduce any speech production difference in genders, the results of gender-dependent classification were higher. Therefore, it might be beneficial, when dealing with prosody features, to separate gender for more accurate depression recognition.

Positive vs. Negative Classification Using Prosody Features

As one of the classification tasks is to differentiate depressed and controls, while expressing positive and negative emotions, two related questions are used from the interview, where I assumed that these elicit those expressions. As explained in Section 3.1, the "Good News" question is used for positive expression, and the "Bad News" questions for negative expression. Although the same number of subjects are used for the expression-dependent classification, the sample size (duration) is substantially smaller. Table 4.12 shows the relevant sounding segments that are used to extract speech prosody features, which is an extract of Table 4.5.

Duration (min)	"Good News" Question		"Bad News" Question	
	Depressed	Control	Depressed	Control
Subjects	30	30	30	30
Subjects' sounding	9.9	6.6	14.9	9.7
Average subjects' sounding	0.3 ($\pm$ 0.2)	0.2 ($\pm$ 0.1)	0.5 ($\pm$ 0.4)	0.3 ($\pm$ 0.2)

Table 4.12: Total duration of subjects' sounding of positive and negative questions from the interview (in minutes) for each group, as well as average duration and standard deviation (in minutes) *(extract from Table 4.5)*

For consistency with the other classification tasks in this work, several feature selection methods are applied on the functional features using a subset of the interview as described earlier. Variable and fixed ETF and PCA, as well as Min-Max feature normalisation are applied in a leave-one-out cross-validation manner on the specific subset; that is, on positive and negative segments individually. Worth noting is that all features and all ETF features listed in Table 4.7 are fixed for the expression-dependent classification as in the general classification task for comparison, as well as to test their generalisation ability on a smaller subset. Expression-dependent classification results are shown in Table 4.13.

Speech style features	"Good News" question		"Bad News" question	
	Number of Features	Average Recall	Number of Features	Average Recall
All speech style features	504	63.3	504	63.3
All ETF	63	68.3	63	66.7
Variable ETF	23-49	55.0	14-35	55.0
Mutual ETF	16	73.3	8	71.7
Fixed PCA variances	53	61.7	53	65.0
98% of PCA variance	53	61.7	53	65.0

Table 4.13: Expression-dependent correct classification results (in %) using speech prosody features

Comparing positive and negative expression classification, using the "Good News" and "Bad News" questions, respectively, the results were not statistically different. In general, the classification results were similar if not slightly higher in the positive expression than the negative one, with the exception of using PCA where the result was slightly higher for the negative expression. Even though the "Good News" segment duration is only two third of that in the "Bad News", the classification results of the "Good News" segments were comparable with the "Bad News" segments.

Moreover, Table 4.13 shows the number of features that are statistically significant to differentiate depressed and controls based on the positive and negative expressions. As can be seen, the number of features of variable and mutual ETF in the positive expression is almost twice the number in the negative expression. With this finding, I conclude that a significant difference between the two subjects groups when expressing a positive emotion and a smaller difference while expressing negative emotions exist.

Consistent with speech style features, the relatively high classification results with less duration, and the high number of features that differentiate depressed from controls while expressing positive emotions could indicate a noticeable difference in positive emotion expression. On the other hand, the lower classification results and the lower number of features that differentiate depressed from controls while expressing negative emotions could indicate a similarity in negative expression. As I concluded with the speech style features, the same applies to the speech prosody features, i.e. positive emotions are expressed less often in depressed subjects, which is in line with Bylsma et al. [2008], and that negative emotions dominate in depressed subjects, which is in line with Ekman and Fridlund [1987] and Ekman [1994].

While using all interview questions in the general classification, the results were statistically higher than in the expression-dependent classifications when using ETF, and statistically lower when using all features and PCA. Given the reduction in sample size (duration) for each observation (subject), the higher results when using all features and PCA obtained by expression-dependent classifications show a strong effect of expressions, especially for positive expressions. As with the speech style classification, it is not clear whether the higher classification results when using ETF for general classification are based on the larger sample size per observation, or based on the combined expressions of both positive and negative emotions.

4.4 Summary

In this chapter, depressed speech characteristics compared to healthy speech were investigated in an interview context. A further investigation of gender differences, as well as positive and negative expression differences between the two groups was also explored. Even though linguistic cues were not investigated in this chapter, para-linguistic cues including speech style and speech prosody features were a rich source of cues to detect depression.

To extract speech features, the audio file has to be segmented to identify speakers and their turns. To obtain as accurate a result as possible, and to reduce error rates of identifying relevant segments, I manually labelled the interviews to separate the subject's speech from the interviewer, as well as the response time, overlapping speech, and other segments. Moreover, the subjects' segments are further analysed using a voice activity detector to identify sounding and silent segments. Several speech style features were extracted from the manual labelling and several prosody features were extracted from the sounding segments.

Both speech style and speech prosody features were analysed statistically. Several speech style features were found to be statistically significant, including response latency in depressed subjects, possibly caused by higher cognitive load, as well as less involvement and less positive reaction in depressed patients. Speech prosody features were analysed statistically as well, where F0, voice quality, energy, shimmer, formants, and MFCC had some of their statistical measures being statistically significant to differentiate depressed from controls. These significant results from prosody features indicated a reduced control of vocal cords in speech production in depressed subjects.

The performance of speech style and speech prosody functional features in the general classification task was relatively higher than in low-level features. Prosody features were also analysed individually. Shimmer and formants consistently gave high classification results regardless of the feature type (i.e. low-level or functionals) and regardless of the classifier used (i.e. GMM, hybrid GMM-SVM, and SVM). For both speech style and prosody features, mutual ETF feature selection resulted in the highest AR results, which supports the hypothesis that the T-statistic threshold can be used for feature selection in a binary classification task.

Gender-dependent classification was also investigated to show gender differences and the generalisation ability of the system with a reduction in the number of observations. In both modalities (speech style and speech prosody), female depression recognition was statistically higher than in males. Moreover, the number of features, found to be statistically different in depressed females from control females, was higher than that in males, which is in line with previous gender differences studies.

Beside gender-dependent classification, expression-dependent classification of positive and negative expressions was investigated. Even though the duration of negative expression segments is longer than the duration of the positive expression segments, positive expression performed higher than negative expression in both speech style and prosody features. The same applies in the number of features that significantly differentiate depressed from controls, being higher in the positive expression than in the negative expression. These findings imply similarity between depressed and controls while expressing negative emotions, and a higher difference between the two groups while expressing positive emotions.

For both gender-dependent and expression-dependent classifications, the sample size and number of observations was reduced substantially. Yet, recognition rates in both speech style and speech prosody modalities remained consistent. That could be explained by the psychological thin-slicing theory.

That is, when using a brief observation (a thin slice) of behaviour, the prediction outcome would be at levels above that expected by chance [Ambady and Rosenthal, 1992] compared to using full observation. Moreover, the thin-slicing theory could be described as the ability to find patterns of behaviour based only on narrow windows of experience. Therefore, in expression-dependent classification, for example, the duration of the used interview is a thin-slice of the duration of the full interview. Yet, the expression-dependent classification performance was comparable to the general classification performance. The same applies to the gender-dependent classification, where the number of observations was reduced, and yet, the gender-dependent classification results were comparable to the mixed-gender classification results. This theory supports the flexibility and robustness of the proposed system to sample size reduction.

Given that the gender-dependent and expression-dependent classification results were comparable, if not higher than general classification results, it might be beneficial to model genders separately, and to include a longer duration of positive expressions to obtain more accurate results, especially from speech modalities.

Since the speech modalities investigated in this chapter (speech style and prosody) were informative and had discriminatory power to detect depression individually, I assume that fusing their features would increase the recognition results. Moreover, investigating and fusing nonverbal cues from the video channel with the speech features might not only increase the accuracy results but also increase the confidence level of the classification results. Therefore, the next chapter investigates nonverbal cues from eye and head modalities, while the chapter after that investigates fusion techniques of the analysed modalities of speech, eye and head on depression recognition, and investigates the influence of each modality in contributing to the final results.

Chapter 5

Depressed Appearance Characteristics

The previous chapter investigated speech characteristics for depression, finding it to be an extensive source of cues to detect depression. However, the body language also holds rich information about the subject's emotions and mood. Section 2.1.2 reviewed nonverbal symptoms of depression including eye activity, head pose, facial expression, and body posture. Out of these modalities, only eye activity and head pose are investigated here for several reasons. Unlike facial expressions to characterise depression, which have been investigated in a number of previous studies, as reviewed in Section 2.3.3, the literature on eye and head modalities is still in its infancy. Also, as the main dataset (BlackDog) does not include full body recordings, body posture and hand movement could not be investigated here.

To address Q2 of the research questions listed in Chapter 1, this chapter uses the BlackDog dataset to extract and analyse eye activity and head movement features for characterising depression. This chapter describes the feature preparation and extraction approach, as well as the analysis of the extracted features from both eye and head movements in Section 5.1 and Section 5.2, respectively.

Consistent with speech modalities, for both eye activity and head pose modalities, several experiments are conducted, aiming to explore the differences between depressed and non-depressed nonverbal behaviour. First the extracted features are analysed statistically, then the extracted features are used for several classification tasks including: comparing low-level with functional feature classification, comparing male with female groups (gender-dependent) classification, and comparing positive and negative emotion expression (expression-dependent) classification.

5.1 Eye Activity

As explained in Section 2.2.3, the region of interest or the object in question has to be identified, located, and tracked (in case of videos) in order to extract the object's features for further analysis and classification. In this section, I am interested in extracting eye activity including iris movements and blinks. The following sections detail the approach used to locate and track the eyes, and explain the extracted

features, which will lead to the statistical analysis and classification tasks.

5.1.1 Locating and Tracking the Eyes

The detection of the human eye is in general a difficult task as the contrast between eye and skin is generally poor. Also, the head anatomy around the eyes may result in poor illumination for eye tracking [Hansen and Pece, 2005]. Blinking could be another difficulty for eye tracking [Morris et al., 2002], as the defined structure for the eye is different while blinking.

Eye tracking using computer vision techniques has the potential to become an important component in computer interfaces. Several devices are often employed for eye-movement measurement including eye-marker cameras, head-mounted corneal reflex illumination, contact lens method, etc. [Young and Sheena, 1975]. In computer vision, techniques for locating the head then finding the eye feature points are used [Hager and Toyama, 1998; Morris et al., 2002]. Two types of acquisition processes are commonly used in eye tracking: passive and active approaches. The passive approach relies on natural light such as the ambient light reflected from the eye. It is often the case that the best feature to track is the contour between the iris and the sclera known as the limbus [Li et al., 2005]. On the other hand, the active illumination approach such as the infrared imaging uses the reflective properties of the pupil when exposed to near infrared light (dark or bright pupil effects). Infrared imaging uses the pupil, rather than the limbus, as the strongest feature contour.

However, none of the previous techniques are applicable to the datasets used in this work, as the recordings were obtained using visible light cameras, which are not specific for eye tracking. Therefore, a technique to locate and track the eyes as moving objects is needed.

The goal is to locate the eye corners, the borders of the eyelids and the iris centre for each eye. Several computer vision techniques have been investigated. The face is located first, and then an approximate eye position is identified. Several corner detection techniques for the eye corners, curve fittings techniques for the eyelids, and iris detection techniques for iris centre are experimented on, none of which where robust to head movement and eye blinks, as well as not accurate enough for the application. Therefore, as I aim to detect and track the eye movements and blinks as accurately as possible, an eye specific model is trained, as per the following description.

Active appearance models are used and trained specifically for the eyes region with different states of the eyes (open, half open, and closed) and with different amounts of head rotation. I tried a different number of points and different shapes for the eye AAM. I started from 5 points for each eye (the corners, the centre of eyelids, and iris centre), which were not enough for the task. While increasing the number of points around the eyes did not increase the accuracy of locating or tracking them, I included the eyebrows, as I believe that they would work as an anchor for the eye and reduce locating and tracking errors. For each eyebrow, I started with 5 points (a point at each end, and the other 3 distributed in the middle of the eyebrow)

and then increased the points to 7, where the model was not accurate in locating and tracking the eyes region. Therefore, I increased the number of points for the eyebrows and changed their location. I used 12 points for each eyebrow, where the points are located at the border of the eyebrow. I believe this configuration works better because of the high contrast between eyebrow and skin, which made these points robust to locate and track with minimum error rate. Inspired by [Bacivarov et al., 2008], the final model contains 74 points for the eyes region including the eyebrows. The location of the points (in order) is shown in Figure 5.1.

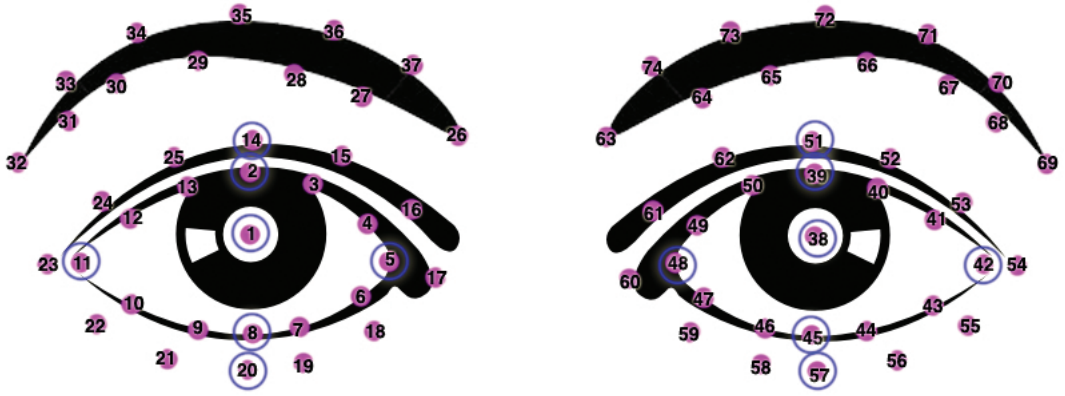
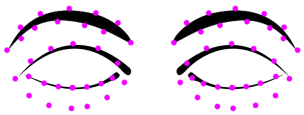


Figure 5.1: Final eye AAM model (for open eye state) with 74 points in the order shown. *(Only marked points are used for feature extraction)*

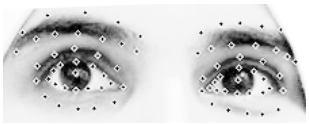
As can be seen from the model, points around the upper and lower eye folds are included in the model to give stability for tracking in the closed eye state. The eye AAM was trained in a subject-specific manner. That is an eye AAM was trained for each subject using images from that subject interview. I used this subject-specific eye AAM to get accurate results, because when I used a generic eye AAM model trained using images from different subjects from the BlackDog interviews, the model fitting and tracking was not accurate. The specification about the training the model is as follows:

On average, 45 images per subject were manually selected from the interviews of the BlackDog dataset. The selection of images was based on different eye status (e.g. open, half open, closed eye) and head position variation. The subject-specific eye AAM was built using linear parameters to update the model in an iterative framework as a discriminative fitting method [Saragih and Goecke, 2006]. Following the annotation and model-building process, the points of the trained model are initialised in a semi-automated manner, i.e. face detection [Viola and Jones, 2001] is performed in the first frame, then the rough eyes' location is estimated using the top third of the detected face. Finally, a fitting function is called. If the fitting function is not accurate enough, the steps of changing the model location and the calls for the fitting function are manually repeated before the tracking is started for the entire video. The trained model fits on and tracks the subject's eyes for the entire interview, producing the positions of the 74 points in each frame. The eye AAM model for open and close

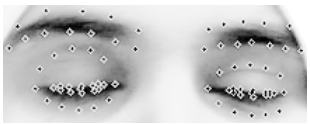
(a) 74 points AAM for Open Eyes



(b) 74 points AAM for Closed Eyes (*eyelid points overlapped)*



(c) Fitting AMM Example for Open Eyes



(d) Fitting AMM Example for Closed/Blinked Eyes

Figure 5.2: Eye AAM with 74 points for both eyes including eyebrows *(showing the overlapping points between upper and lower eyelids including iris centre for closed eyes)*

states is illustrated in Figure 5.2 (a)&(b). An example of the fitted AAM model is also shown in Figure 5.2 (c) & (d).

5.1.2 Eye Activity Feature Extraction

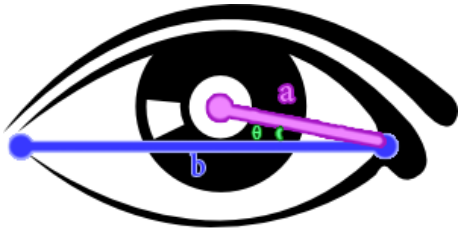
The goal is to extract iris movement features (e.g up/down and left/right), as well as eyelid distance features (e.g. wide open, blinks). From these features, statistical measures were also calculated, as well as including duration features, such as the duration of gazing in one direction and the duration of blinks, which will be elaborated later. To calculate these features, only 7 points of each eye were used from the eye AAM model (left eye points: 1, 2, 5, 8, 11, 14, 20; right eye points: 38, 39, 42, 45, 48, 51, 57, as marked in Figure 5.1).

Eye Activity Low-level Features

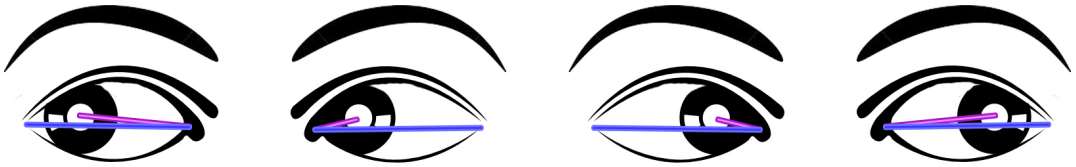
For each eye, three features were extracted: iris horizontal movement, iris vertical movement, and eyelid distance.

The horizontal movement: is measured by the length of the line connecting the inner corner of the eye and the iris centre. The horizontal movement has been normalised based on the line connecting the eye corners (see Figure 5.3). The normalising procedure is to reduce variability of the head distance from the camera and variability of different eyes region shape and size. The longer the line, the further the iris is from the inner eye corner and vice versa.

The vertical movement: is measured by the angle between the previous two lines that measure the horizontal movement (see Figure 5.3). A larger angle indicates a higher iris position (i.e. looking up) and vice versa.

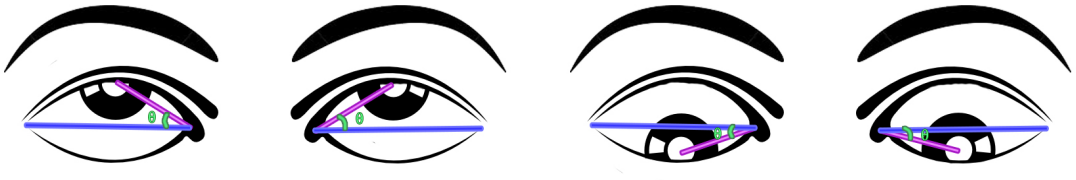


Left-right *(length of a ÷ length of b)*
and up-down (θ) Movement features



Example of looking to the subject's right

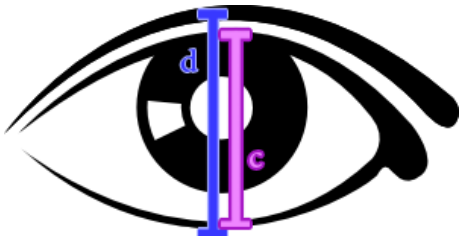
Example of looking to the subject's left



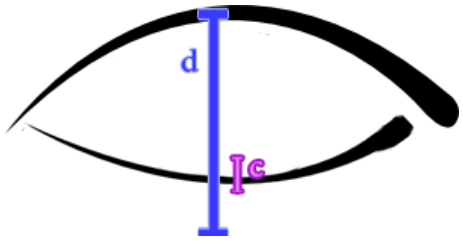
Example of looking up

Example of looking down

Figure 5.3: Extracting and normalising eye movement features



Open Eye Eyelid Distance
(length of c ÷ length of d)



Closed Eye Eyelid Distance
(length of c ÷ length of d)

Figure 5.4: Extracting and normalising eyelid distance and blink features

Distance between the eyelids is measured by the length of the line connecting the centre of both eyelids, normalised by the line connecting the centre of the eye fold and the centre of the outer border of the lower eyelid, which also applies for closed eyes (see Figure 5.4).

These three features of each eye are extracted for each frame (30 frames per second), then their velocity and acceleration are extracted to give a total 18 features per

frame. Outlier frames, which are caused by erroneous fitting of the eye AAM, or the absence of the eyes in a frame, are detected using Grubbs' test for outliers [Grubbs, 1969]. The detected outlier frames are skipped (an average of 1.8% of the frames were skipped per interview).

Eye Activity Functional Features

Moreover, a total of 126 statistical features "functionals" were extracted, which are:

- Maximum, minimum, mean, variance, and standard deviation for all 18 low-level features mentioned earlier (5×18 features)
- Maximum, minimum, and average of: duration of looking left, right, up and down, as well as blink duration for each eye (3×2 eyes $\times$ 5 features)
- Closed eye duration rate, and closed eye to open eye duration rate for both eyes (2 eyes $\times$ 2 features)
- Blinking rate for both eyes (2 eyes $\times$ 1 feature)

A blink is detected when the normalised eyelid distance is lower than the average of the normalised eyelid distance minus half the standard deviation of that feature for each subject. The same applies for the gaze direction, where a gaze direction is detected when the normalised feature (i.e. distance or angle from inner eye corner to iris centre) is lower than the average of that feature plus/minus half the standard deviation of that feature for each subject.

5.1.3 Statistical Analysis of Eye Features

Table 5.1 shows only the 21 features that have a significant T-statistic result (i.e. $p < 0.05$) out of all 126 extracted functional features. The state of the T-test specifies the direction of effect. That is, identifying which group is higher than the other in a particular feature.

The significant features can be categorised into four categories: movement, looking direction duration, distance between eyelids, and blinks.

Movement features are calculated over the low-level features, to extract the mean, maximum, minimum, etc. Only two features are statistically significant between the two groups (depressed/control), which are the mean and the minimum of right-left movement. Given that the interviewer was approximately in middle of the interviewee's field of view, the average distance between the subject's right eye inner corner and the iris centre (i.e. right-left movement) is larger in depressed subjects, meaning more gazing to the right, which might be an indicator for avoiding eye contact with the interviewer. On the other hand, the left eye has no significant result for the same feature.

This might be due to, when a person looks to the right, as illustrated in the examples in Figure 5.3, the variation of the distance between the subjectâĂŹs left eye

Eye Activity Feature		Direction
Left Eye	Mean right-left movement	D>C
	Minimum right-left movement	D>C
	Mean eyelid distance	C>D
	Closed rate	D>C
	Closed to open eye rate	D>C
	Maximum blink duration	D>C
	Mean blink duration	D>C
	Maximum looking down duration	D>C
	Minimum looking up duration	C>D
	Maximum looking right duration	D>C
Right Eye	Mean eyelid distance	C>D
	Minimum eyelid distance	C>D
	Standard deviation eyelid distance	D>C
	Variance eyelid distance	D>C
	Closed rate	D>C
	Closed to open eye rate	D>C
	Maximum blink duration	D>C
	Mean blink duration	D>C
	Maximum looking up duration	D>C
	Rate of looking right	D>C
	Maximum looking right duration	D>C

Table 5.1: Gender-independent significant T-test results of eye movement features for the interview *(Direction of effect is reported to show which group depressed (D) or controls (C) is higher than the other in the analysed feature)*

inner corner and the iris centre is small, while the variation of the distance between the subjectâĂŹs right eye inner corner to the iris centre is larger. Therefore, subtle changes in the small variation might not be detected as accurately as larger changes by the tracker.

For looking direction duration features, a few features were significantly different between the two groups. Looking to the right duration (maximum and rate) was significantly longer for the depressed group, which could indicate eye contact avoidance. Worth noting is that the door of the recording environment was located to the subject's right. If it is assumed that depressed subjects were looking longer at the door, that could show conversational avoidance or subconsciously finding a way out. Studies on infants and autistic children have shown that children stare at the door when they are upset or bored (e.g. [Eisenberg and Spinrad, 2004; Grandin, 1995]), while no similar studies could be found on adult subjects. Therefore, such conclusion for depressed subjects needs more analysis with a larger dataset. Moreover, looking down duration was also longer in depressed subject, which is also a common reaction associated with sadness.

Interestingly, the results show that the average distance between the eyelids is significantly smaller in depressed subjects. The smaller distance between the eyelids in depressed subjects might be an indication of fatigue, which is a common symptom of depression.

Also, even though the blink rate was not significantly different between depressed and control subjects, the results show that the blink duration was significantly longer in depressed subjects. The latter finding suggests that, assuming that the physical

need of a blink to lubricate the eye is similar in both groups, the longer duration of a blink could indicate eye contact avoidance and fatigue, which are common symptoms of depression.

5.1.4 Classification Using Eye Activity Features

Several classification tasks were performed for comparison. First, the performance of low-level and functional features was compared in the classification task. Then, gender-dependent classification was investigated. Finally, the classification results of different emotion expressions (positive and negative) were compared.

For all classification tasks, even though the extracted low-level features are pre-normalised (as explained in Section 5.1.2), the functional features need further normalisation to be on a unified scale. Therefore, functional features are normalised using Min-Max normalisation in a leave-one-out cross-validation manner.

Low-level vs. Functional Eye Activity Features Classification

Feature Type	Classifier	Features	Number of Features	Depressed Recall (Sensitivity)	Control Recall (Specificity)	Average Recall
Low-level	GMM (21 mixtures)	frame-by-frame	18	50.0	63.3	56.7
	GMM+SVM	GMM model	21 mixtures	73.3	83.3	78.3
Functional	SVM	All features	126	66.7	70.0	68.3
		All ETF	21	76.7	86.7	81.7
		Variable ETF	16-23	73.3	73.3	73.3
		Mutual ETF	13	83.3	86.7	85.0
		Fixed PCA variances	41	66.7	73.3	70.0
		98% of PCA variances	41-42	66.7	73.3	70.0

Table 5.2: Correct classification results (in %) of eye activity low-level and functional features

Although eye movements would most likely be used as a complementary cue in a multimodal affective sensing system for depression, rather than as the sole measure, they performed well on their own.

Table 5.2 shows the classification results in term of average recall for both frame-by-frame features (low-level) and functionals using different classifiers and feature selection. As explained in Section 3.1, GMM can deal with low-level data of different duration (different number of frames), which is the reason it has been used for low-level features. SVM requires an equal length of feature vector for all observations, which is suitable for functional features. Since the GMM models have equal feature vector length (i.e. number of mixtures) for each subject, SVM could use the GMM models for each subject as observations. Therefore, I experimented on using a hybrid classification using GMM models as features for an SVM classifier.

For low-level features, two methods were performed: using GMM as a classifier and then as features for SVM as explained in Section 3.1. As can be seen in Table 5.2, the GMM alone gave lower results. The GMM results are shown here as a baseline for comparison with the other classifications. Low-level features gave 79% AR using

the hybrid classifier of GMM models with SVM, which is statistically above chance level, and also statistically comparable to using functional features.

For functional features, several feature selection methods were performed as listed in Section 3.1. Generally, the performance of functional features is high, giving on average 75% AR, which is statistically above chance level. However, the lowest classification result was obtained when using all 126 extracted functional features (68% AR). Using PCA, was similar to using all features.

The T-statistic threshold was also used as filtering method, where features that exceeded a statistical threshold set in advance by a t-value corresponding to an uncorrected p-value of 0.05 ($p < 0.05$) (I refer to these features as ETF) are selected in three approaches as described in 3.1, which are: using all ETF listed in Table 5.1, using variable ETF selected based on cross-validation turns, and using mutual ETF selected based on cross-validation turns.

The goal of using all ETF listed in Table 5.1 is to have a fixed list of features to compare with different classification tasks, even though the list is calculated over the entire interviews using all subjects regardless of their gender. Using all ETF gave a high accuracy compared to other methods.

Variable ETF are the features that have been selected based on the training set and applied to the testing set in each cross-validation turn, regardless of the fact that the features combination and number are variable in each turn. On the other hand, the mutual ETF only select the features that intersect in each turn. The highest performance was obtained with mutual ETF where the result was 85% AR. While the variable ETF gave high result, it was considerably lower than using the mutual ETF, which is expected as the selection of mutual ETF is what work best with all cross-validation turns.

PCA was also used to reduce dimensionality and select the most promising PC variances (98% in this case). As the 98% of PC variances might differ between turns, the number of PCs was also fixed, so that the amount of variance represented was always at lest 98%, to ensure fair comparison between cross-validation turns (as explained in Section 3.1). Nevertheless, both PCA approaches performed equally in terms of classification results. The similarity of the performance of PCA approaches is caused by the selected PC variances in both approaches being almost similar.

Regardless of the feature selection method used for the eye modality, the high classification results support previous studies, which concluded abnormality in patients' ocular motor system [Lipton et al., 1980; Abel et al., 1991; Crawford et al., 1995; Kupfer and Foster, 1972]. Using only the features that exceeded the t-statistic in all three approaches performed remarkably better than using all features. Even with the large reduction of the number of features (PCs) in PCA compared with all features, the classification result was not statistically different to using all features.

Table 5.2 also shows the classification recall for both depressed and control groups. With all classification and feature comparisons, control recall is higher than depressed recall. That indicates that misclassification in the depressed group is higher than the misclassification in the control group. Performing error analysis, none of the meta-data including depression score, age, medication seem to be the

reason behind the classification errors. A more detailed error analysis is presented in Section 6.3.

Overall, functional features performed better than the low-level features. Using GMM models as features for SVM performed similarly to using all ETF features, which might indicate that the GMM model captures most informative features in low-level data. However, selecting the number of mixture for GMM is not a trivial task, as the empirical selection of GMM mixtures is time consuming and not practical.

Since functional features performed better than low-level features, further classification tasks will use functional features only, to reduce confusion and increase the emphasis in the goal of the comparison.

Gender-dependent Classification Using Eye Activity Functional Features

To investigate the effect of gender in detecting depression, men and women were manually separated in each group (depressed/control). I acknowledge the large reduction of sample size as well as the number of observations. However, such investigation could give an insight not only into gender differences but also into the proposed system's flexibility in the case of sample reduction. The duration and number of subjects of each gender in each group are shown in Table 5.3.

Duration (min)	Male Subjects		Female Subjects	
	Depressed	Control	Depressed	Control
Subjects	15	15	15	15
Total	123.8	112.0	182.5	87.8
Average	8.3 ($\pm$ 3.7)	7.5 ($\pm$ 2.0)	12.1 ($\pm$ 6.4)	5.6 ($\pm$ 1.8)

Table 5.3: Total duration of the entire interview (in minutes) for each gender in each group, as well as average duration and standard deviation (in minutes)

To ensure a fair comparison with previous gender-independent classification, the feature that had a significance level based on T-test are kept the same in this gender-dependent classification task as listed in Table 5.1. The PCA was performed in a leave-one-out cross-validation manner, for each gender-dependent group individually. Moreover, the Min-Max normalisation method is also performed in a leave-one-out cross-validation manner, for each gender-dependent group individually. The results of the gender-dependent classification are shown in Table 5.4.

Functional features	Males		Females	
	Number of Features	Average Recall	Number of Features	Average Recall
All functional features	126	63.3	126	83.3
All ETF	21	76.7	21	83.3
Variable ETF	5-16	63.3	13-18	86.7
Mutual ETF	5	80.0	13	86.7
Fixed PCA	24	63.3	24	80.0
98% of PCA variance	24	63.3	24	80.0

Table 5.4: Gender-dependent correct classification results (in %) using eye activity functional features

Regardless of the features used for classification, the results show a higher recognition rate of depression in females than males. With the exception of using all and mutual ETF, the recognition rate has considerable differences between males and females classification. The studies about gender differences in depression nonverbal behaviour pointed out that depressed women are more likely to be detected than depressed men [Nolen-Hoeksema, 1987; Troisi and Moles, 1999; Stratou et al., 2013]. The previous studies supports the finding in the differences between genders when depressed.

Using all and mutual ETF performed statistically better than other feature selection in males, while the performance in female classification results with all different method of feature selection were consistently high. While all ETF were selected based on the entire interviews using all subjects, applying the same feature set on both gender-dependent classifications resulted in high ARs. This finding indicates that these features generalise even on a smaller subset.

Expression-dependent Classification Using Eye Activity Functional Features

Investigating positive and a negative expression, two related questions were used from the interview that were assumed to elicit the expressions in question, as described in Section 3.1. The duration and number of subjects in each group with each selected question are shown in Table 5.5.

Duration (min)	"Good News" Question		"Bad News" Question	
	Depressed	Control	Depressed	Control
Subjects	30	30	30	30
Total	34.1	21.3	36.1	28.7
Average	1.1 ($\pm$ 0.7)	0.7 ($\pm$ 0.4)	1.2 ($\pm$ 0.9)	1.0 ($\pm$ 0.6)

Table 5.5: Total duration of positive and negative questions from the interview (in minutes) for each group, as well as average duration and standard deviation (in minutes)

As with the gender-dependent classification, to ensure a fair comparison with previous classifications, the features that exceeded the t-statistic are kept the same as listed in Table 5.1. The Min-Max feature normalisation and PCA were performed in a leave-one-out cross-validation manner, for each expression individually. The results of the expression-dependent classification are shown in Table 5.6.

For the "Good News" question (positive expression), recognising depression had high accuracy results that were considerably above chance level, and were almost as accurate as when using the entire interviews, and even remarkably better for using all features and PCA (compare Table 5.2 and Table 5.6 "Good News"). Getting good recognition rates from such a small sample indicates the clearly noticeable differences in expressing positive emotions in depressed and control subjects.

On the other hand, in most cases analysing the "Bad News" question (negative emotion), gave worse recognition rates than using the entire interviews or the positive question (compare Table 5.2 and Table 5.6). This indicates that both groups

Functional features	"Good News" question		"Bad News" question	
	Number of Features	Average Recall	Number of Features	Average Recall
All functional features	126	75.0	126	66.7
All ETF	21	75.0	21	68.3
Variable ETF	14-24	65.0	7-12	73.3
Mutual ETF	13	71.7	7	78.3
Fixed PCA variances	35	78.3	35	68.3
98% of PCA variance	35-36	80.0	35-37	68.3

Table 5.6: Expression-dependent correct classification results (in %) using eye activity functional features

express negative emotions in a similar manner. This finding supports the previous finding from investigating speech style and speech prosody regarding positive emotions being expressed less often in depressed subjects at all times [Bylsma et al., 2008] and, hence, negative emotional eye behaviour has less discriminatory power than for positive emotions.

Once more, the classification results from using all ETF selected based on the entire interview were high in both positive and negative subset. This finding supports the findings with the gender-dependent classification results that these features have the ability to generalise on smaller subsets.

In all classification tasks, the highest classification results were obtained from using mutual ETF, which is expected. Regardless of the drawbacks of such method of feature selection, it produces an equal feature set and feature vector length in each cross-validation turn, which ensures a fair comparison. Moreover, as the cross-validation models are optimised to have equal parameters, it is favorable to have similar feature set as well.

The fixed and variable PCA variance selection did not have a remarkable difference. That might be caused by the range of PCs to be selected not being largely different between cross-validation turns. Nevertheless, the PCA feature selection method performed similarly to using all features if not statistically better, which may support the view that the selected variances represented the feature space well.

Both gender-dependent and emotion-dependent investigations gave relatively good recognition rates, despite the large reduction in the sample size and observations. This result could be explained by the thin-slicing theory, i.e. when using different smaller parts of the interview, the performance is similar if not better than using the entire interview (as explained in Section 4.4). This finding supports the view that the eye activity are informative to classify depression. It also supports the flexibility and robustness of the proposed system to sample size reduction.

5.2 Head Pose and Movement

Even with the little research on head pose and movement for depressed patients as described in 2.1.2, strong indicators for depressed behaviour and conversational style have been shown. In this section, the head pose and movement for depressed in comparison with controls are investigated, and their discriminative power for detecting

depression.

5.2.1 Locating the Face to Estimate Head Pose

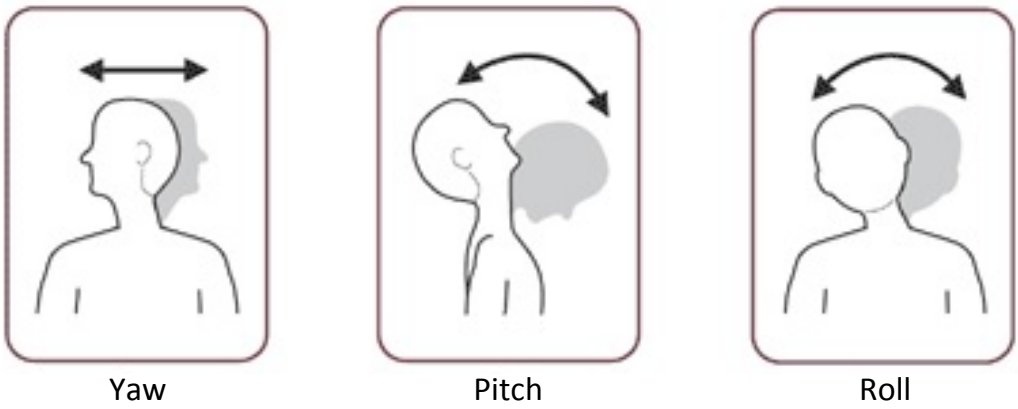


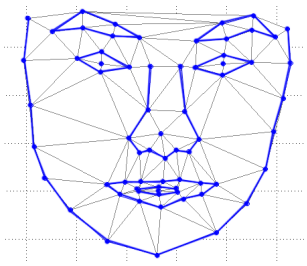
Figure 5.5: Head rotation angles: Yaw, pitch and roll [NeuroCom, 2014]

To estimate the head pose and movement behaviour, the face or some facial landmarks should be located and tracked, then a 3-degree of freedom (DOF) head pose could be calculated (see Figure 5.5). Even though the face has a defined structure and, in the typical videos of datasets available, occupy a reasonably large space in each frame, it might not be trivial to locate the face accurately. Automatic head pose estimation using computer vision techniques have been surveyed in the Murphy-Chutorian and Trivedi [2009] study. Such methods include template matching, rule-based, motion-based, and deformable models.

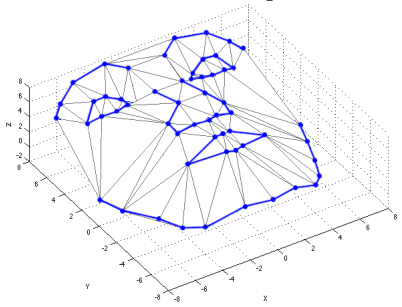
One method of head pose estimation is by determining the geometry between local features, such as the eyes, mouth, and nose tip. For example, Nikolaidis and Pitas [2000] located 3 points in the face (iris centres and mouth centre). Horprasert et al. [1997] and Gee and Cipolla [1994] located 5 points of different landmarks for each study, while Wang and Sung [2007] located 6 points. All studies then calculate an estimation for the head pose. Another technique approximated the head pose using the shape of the head by a 3D cylinder [Seo, 2004].

On the other hand, deformable models, including the AAM, where specific facial points are labelled and trained to create a 2D model, then the head pose is estimated using the direction of the first principal component of the principal components analysis [Murphy-Chutorian and Trivedi, 2009]. Moreover, Martins and Batista [2008] estimated 3D head pose by combining AAM with an algorithm called Pose from Orthography and Scaling with ITerations (POSIT) [Dementhon and Davis, 1995], which will be explained later on.

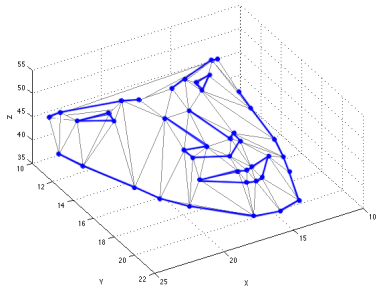
However, not all methods could estimate all three dimensions of the head pose (pitch, yaw and roll as in Figure 5.5). Some of the methods are limited to a small range of head orientations. Also, when using a small number of facial points, some of the methods will not be robust to partial occlusion, and the head pose estimation



(a) 2D AAM face model using 68 landmarks



(b) 3D face model using 58 landmarks



(c) The 3D Model projected onto the 2D Model
Using Only Matched 46 Points

Figure 5.6: 3D to 2D AAM projection for estimating the head pose

will be affected by mistakenly locating any point. In this work, the POSIT algorithm is used to estimate the head pose using several facial points (46 points). POSIT is an algorithm that detects and matches at least 4 points in a 2D image to a defined 3D object, then finds the geometry of the 3D object (orientation and translation) [Dementhon and Davis, 1995]. Once the estimated orientation is calculated, the pose of the object could be extracted. POSIT is a useful alternative to popular pose algorithms because it does not require an initial pose estimation, and because it is easy to implement as well as faster to run [Dementhon and Davis, 1995].

For the BlackDog dataset, an average of 30 images were automatically selected (almost every 250 frames) per subject having different head position variation. These images were annotated using 68 points (see Figure 5.6 (a)). These annotated images were used to build subject-specific face AAMs, using linear parameters to update the model in an iterative framework as a discriminative fitting method [Saragih and

Goecke, 2006]. The points of the trained model were initialised in a semi-automated manner, i.e. face detection [Viola and Jones, 2001] was performed in the first frame, then a fitting function was called. If the fitting function was not accurate enough, the steps of changing the model location and the calls for the fitting function were manually repeated before the tracking was started for the entire video. The trained model fits on and tracks the subject's face for the entire interview, producing the position of the 68 landmarks in each frame.

To obtain the 3D pose of the subject's head, a 3D face model is projected onto the 2D AAM facial points. For the 3D model, a 58-points 3D face statistical anthropometric model was used [Martins and Batista, 2008] (see Figure 5.6 (b)). Since the 2D AAM uses 68 points and the 3D model used 58 points, only the 46 points that correspond in position for both models (see Figure 5.6 (c)) were chosen. The resulting 46 points of the 3D model are projected on to the acquired tracked 46 points of the 2D AAM to estimate the head pose using the POSIT algorithm. POS assumes the image was obtained by a scaled orthographic projection. POSIT adds multiple iterations to POS, that is, the rotation and scale matrices are re-estimated until no improvement to the pose projection is detected.

5.2.2 Head Pose and Movement Feature Extraction

Once the head pose is estimated, 3-DOF are extracted from the rotation matrix: yaw, roll, and pitch in each frame, then several functional features based on statistical measurements over time are calculated.

Head Pose and Movement Low-level Features

The output of the POSIT algorithm are the rotation matrix and a scale vector. The 3-DOF are then calculated from the rotation matrix as follows:

Defining the rotation matrix R as:

$$R = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} \tag{5.1}$$

rotation angles could be extracted as follows:

$$Yaw = \tan^{-1}(r_{21}/r_{11}) \tag{5.2}$$

$$Roll = \tan^{-1}(r_{31}/\sqrt{r_{32}^2 + r_{33}^2}) \tag{5.3}$$

$$Pitch = \tan^{-1}(r_{32}/r_{33}) \tag{5.4}$$

These three DOF are extracted for each frame (30 fps), then their velocity and acceleration are extracted to give a total of 9 features per frame. Outlier frames, which are caused by incorrect fitting of the AAM or failing to converge to a 3D model, are detected using Grubbs' test for outliers [Grubbs, 1969], these detected outlier frames

are skipped (an average of 3.5% of the frames are skipped per interview).

Head Pose and Movement Functional Features

Over the duration of each subject's interview, a total of 184 statistical features ("functionals") were extracted, which are:

- Maximum, minimum, range, mean, variance, and standard deviation for all 9 low-level features mentioned earlier. (6 $\times$ 9 features)
- Maximum, minimum, range and average duration of: head direction left, right, up and down, tilting clockwise and anticlockwise. (4 $\times$ 6 features)
- Head direction duration rate, and rate of different head directions for non-frontal head direction for all directions mentioned above. (2 $\times$ 6 features)
- Change head direction rate for all directions mentioned above. (1 $\times$ 6 features)
- Total number of changes of head direction for yaw, roll, pitch, and all directions. (1 $\times$ 4 features)
- Maximum, minimum, range, mean, variance, duration, and rate for slow, fast, steady, and continuous movement of yaw, roll, pitch. (7 $\times$ 3 DOF $\times$ 4 features)

The above duration features are detected when the feature in question is higher than a threshold. The threshold is the average of the feature in question plus the standard deviation of that feature for each subject.

5.2.3 Statistical Analysis of Head Features

Head Pose Feature	Direction
Maximum yaw	C>D
Standard deviation of yaw velocity	C>D
Maximum pitch velocity	C>D
Maximum pitch acceleration	C>D
Range of pitch acceleration	C>D
Range duration of looking right	D>C
Average count of looking left	C>D
Average count of looking right	C>D
Average count of tilting anticlockwise	C>D
Average count of looking down	C>D
Minimum count of steady yaw movement	C>D
Average duration of steady yaw movement	C>D
Average duration of continues yaw movement	C>D
Average duration of continues roll movement	C>D

Table 5.7: Gender-independent significant T-test results of head pose and movement features *(Direction of effect is reported to show which group depressed (D) or controls (C) is higher than the other in the analysed feature)*

Table 5.7 shows only the features that have a significant p-value based on the T-statistic results of analysing the functionals extracted from the interviews. The significant features could be divided into four categories: movement angles, movement speed, looking direction, and movement type.

None of the basic movement angles were significantly different between depressed and control groups, with the exception of the maximum yaw angle, being higher in controls. Such feature could indicate a higher range of head poses for controls compared with depressed.

For movement speed, such as velocity and acceleration of certain angles, velocity and acceleration of pitch and yaw were faster in controls. These features display that depressed patients move their head more slowly than controls. Slower movements for depressed patients are expected as an indication of fatigue, which is a common symptom of depression [Prendergast, 2006].

As with eye activity features, head looking direction duration and count were analysed. Given that the interviewer approximately stands in a centred position in front of the subject, the maximum and the range duration of looking to the right were longer in depressed subjects, which might be an indicator of avoiding eye contact with the interviewer [Fossi et al., 1984]. The average number of times healthy controls move their head left to right and roll their head clockwise to anticlockwise, is higher than for depressed patients. This finding indicates that depression sufferers' head movements are significantly reduced compared to healthy controls, which is in line with [Hale III et al., 1997; Waxer, 1974]. It has been also found that the average duration of looking down is longer in depressed individuals, which could be an indicator of avoiding eye contact with the interviewer [Fossi et al., 1984; Waxer, 1974].

Movement type, such as steady head pose and continues change of head pose were extracted and analysed. Steady head movements from left to right, on average is longer in duration with depressed patients. At the same time, continuous (faster) head movements left to right and for the roll clockwise to anticlockwise, the average duration is longer in healthy controls. These findings are another indication of having less and slow overall head movements in depressed subjects. I conclude that head movements in general are significantly different between depressed patients and healthy subjects due to psychomotor retardation [Parker and Hadzi-Pavlovic, 1996; Parker et al., 1994].

5.2.4 Classification Using Head Pose Features

Similar to the classification tasks performed on eye activity features, automatic classification of depression from both frame-by-frame features and functionals of head pose and movement is performed and reported in this section. Although the head pose and movement would be used as a complementary cue in detecting depression in practice, their recognition rates on their own could show whether they hold effective cues in diagnosing depression. Gender-dependent and expression-dependent classification task are also investigated, to give a deeper insight of the head pose and movement pattern in both depressed and control groups.

Low-level vs. Functional Head Pose Features Classification

Feature Type	Classifier	Features	Number of Features	Depressed Recall (Sensitivity)	Control Recall (Specificity)	Average Recall
Low-level	GMM (7 mixtures)	frame-by-frame	9	53.3	53.3	53.3
	GMM+SVM	GMM model	7 mixtures	70.0	66.7	68.3
Functional	SVM	All features	184	63.3	70.0	66.7
		All ETF	14	70.0	76.7	73.3
		Variable ETF	8-15	76.7	50.0	63.3
		Mutual ETF	7	80.0	70.0	75.0
		Fixed PCA	40	60.0	63.3	61.7
		98% of PCA variance	40-42	66.7	70.0	68.3

Table 5.8: Correct classification results (in %) of head pose low-level and functional features

Classification results from low-level and functional features using different classifiers and feature selection are shown in Table 5.8. For low-level features, GMM was used as classifier and then as feature clustering. For comparison, the number of mixtures was empirically chosen for the best results for both GMM and GMM with SVM classification. The performance of GMM as a classifier was at chance level (53% AR). However, using GMM models as features for SVM performed statistically above chance level (68% AR).

For functional features, SVM was used for classification using different feature selection methods. Even though all classification results from functional features were above chance level, using all ETF listed in Table 5.7 and mutual ETF performed the highest (73% and 75% AR, respectively) in recognising depression.

Unlike with the eye modality, 98% of PCA variances performed statistically higher than fixed PCA, which is similar to using all functional features. The differences between the fixed and variable PCA variances might be due to the higher range of variability compared to the eye modality PCA variance range. However, the similarity between variable PCA and using all features might indicate that the PCA captured the PCs that highly represent the feature space.

Comparing low-level and functional features, low-level features performance was up to 68% AR using the hybrid classifier (GMM models with SVM), while functionals gave up to 75% AR, which were both above chance level. Even though GMM models with SVM gaves recognition rate similar to using all features and PCA, the empirical selection of the number of mixtures for GMM is a barrier for using this method. Nevertheless, in general, the recognition rate for functional features is substantially higher than the low-level feature results, with all ETF and mutual ETF being the highest. This result implies that the selection of features that exceed the t-statistic is a useful method for reducing dimensionality and selecting feature.

Since the functional feature performance is higher as well as the implementation is easier and more practical than low-level features, the further classification tasks will only report on the classification results from functional features and related feature selection methods.

Gender-dependent Classification Using Head Pose Functional Features

As explored for the eye modalities, gender-dependent classification performance was investigated from head pose features. The same subjects were used, and the same interview segments and duration in order to facilitate a fair comparison, as described in Table 5.3. The gender-dependent classification results are presented in Table 5.9.

	Males		Females	
Functional features	Number of Features	Average Recall	Number of Features	Average Recall
All functional features	184	70.0	184	70.0
All ETF	14	76.7	14	73.3
Variable ETF	5-7	76.7	6-13	90.0
Mutual ETF	5	76.7	5	83.3
Fixed PCA	23	63.3	24	70.0
98% of PCA variance	23-24	70.0	24	70.0

Table 5.9: Gender-dependent correct classification results (in %) using head pose functional features

Even with the large reduction in the number of observations, the recognition rate for the gender-dependent case is statistically similar if not better than the gender-independent classification results. As mentioned previously, the gender differences studies in nonverbal behavior in depressed subjects reported that depressed women are more likely to be detected than depressed men [Nolen-Hoeksema, 1987; Troisi and Moles, 1999; Stratou et al., 2013]. With the exception of variable ETF, where the females classification result was statistically higher the male classification result, the comparison between the gender-dependent classification results showed that on average they are statistically similar. Nevertheless, on average, the depression recognition rate in females (76% AR) is slightly higher than in males (72% AR). Although the head pose and movement classification results from each gender group do not support nor conflict with the conclusions of gender differences studies [Nolen-Hoeksema, 1987; Troisi and Moles, 1999; Stratou et al., 2013], the high performance of head pose features might indicate a physical condition (e.g. fatigue) rather than a behavioural one (i.e. not gender-related). Moreover, this results is in line with the Stratou et al. [2013] study, where the head rotation in depressed male and female subjects was compared, and no difference between the two genders in head rotation was found.

Expression-dependent Classification Using Head Pose Functional Features

In addition to gender-dependent classification, expression-dependent classification is also investigated, similarly to the eye modality (see Section 5.1.4). Moreover, the segments of the interview and their duration were shown previously in Table 5.5. The depression recognition rates based on expression are reported in Table 5.10, where functional features and several feature selection methods were used.

While acknowledging the potential impact of the large reduction of training data from using all interview questions to using only two questions, the differences in expressing positive and negative emotions between depressed and control subjects

Functional features	"Good News" question		"Bad News" question	
	Number of Features	Average Recall	Number of Features	Average Recall
All functional features	184	50.0	184	50.0
All ETF	14	58.3	14	60.0
Variable ETF	13-18	73.3	3-9	65.0
Mutual ETF	11	73.3	2	60.0
Fixed PCA variances	38	73.3	39	63.3
98% of PCA variances	38-42	78.3	39-40	63.3

Table 5.10: Expression-dependent correct classification results (in %) using head pose functional features

were investigated. This was done by evaluating the "Good News" and "Bad News" questions from the interview as explained in Section 3.1.

For the "Good News" question (positive emotion), in general, recognising depression was almost as accurate as when using all interview questions. Getting good recognition rates from such a small subset indicates the clearly noticeable differences in expressing positive emotions in depressed and controls subjects.

On the other hand, analysing the "Bad News" question (negative emotion), gave lower recognition rates than using all interview questions and the positive question. As found with the eye modality, this indicates that both groups express negative emotions in the same or a similar manner. This finding supports the previous finding with the eye modality that positive emotions are expressed less often in depressed subjects [Bylsma et al., 2008]. Moreover, the finding supports the conclusion that negative emotions dominate in depressed subjects [Ekman, 1994; Ekman and Fridlund, 1987] and, hence, negative emotions have less discriminatory power than positive emotions in detecting depression from healthy and depressed subjects.

For the head modality, using all ETF features listed in Table 5.7 for the gender-dependent subset performed well, which was not the case for the expression-dependent subset. That might be due to the selected ETF for head movement not being observed in such small segments of the interview. As found with the eye modality, mutual ETF had the highest recognition rate of depression in all classification tasks using head movement features.

Moreover, like the eye modality, fixed and variable PCA did not have a statistical difference in most cases compared to using all features in the head modality features. Nevertheless, the variable PCA performed slightly better than the fixed PCA, which might indicate that fixing the number of PCs (possibly at a number too low) reduced the probability to capture all the PCs that highly represent the feature space. The alternative variable PCA approach fixes the amount of variance but results in a variable number of PCs. This finding might imply that a better representation of features is acquired when using 98% PCA of variances in each cross-validation turn, and using less variance on some turns affects the representation of the feature space.

Consistent with the eye modality, the large reduction in the sample size, for both gender-dependent and emotion-dependent investigations gave relatively good recognition rates. This result could be explained by the thin-slicing theory [Ambady and Rosenthal, 1992] (see Section 5.1.4).

5.3 Summary

In this chapter, eye movement and head pose patterns were investigated for their discriminative power for recognising depression from video data of subject interviews, as well as whether initial manual gender splitting, and emotion expression influence the recognition rate. Low-level and functional features of both modalities were extracted to be analysed statistically, as well as using different classifiers and feature selection methods. Although both eye activity and head pose features are complementary features that could be fused with other cues (e.g. facial expressions, speech, etc.), they generally gave relatively high recognition results on their own (up to 85% AR for eye activity modality, and up to 75% AR for head pose modality).

To extract eye and head activity features, the eyes and the face have to be located and tracked accurately. For locating and tracking the eyes, a subject-specific eye AAM was built with 74 points for the eyes and eyebrows region. From these points, looking directions and distance between eyelids were extracted and normalised. For locating the face, a face AAM with 68 points was also built. A subset of these points was projected on to a 3D face model to estimate the head pose, from which looking direction and head movements were extracted.

For the eye modality, analysing functional features statistically found that the average vertical distance between the eyelids was significantly smaller in between blinks and the average duration of blinks was significantly longer in depressed subjects, which might be an indication of fatigue and eye contact avoidance. In general, I conclude that eye movement abnormality is a physical cue as well as a behavioural one, which is in line with the psychology literature in that depression leads to psychomotor retardation.

Statistical analyses on head pose and movement behavioural patterns found several distinguishing features: (1) depression sufferers had slower head movements, which may indicate fatigue and/or psychomotor retardation, (2) the duration of looking to the right was longer in depressed patients, which might be an indicator of avoiding eye contact with the interviewer, (3) that overall change of head position in depressed subjects was significantly reduced compared to the healthy controls, and (4) the average duration of looking down was longer in depressed individuals, which could be an indicator of avoiding eye contact with the interviewer.

For both eye and head modalities, depression was classified using low-level and functional features. Modelling the low-level features using GMM and then using the GMM model of each subject as an observation for an SVM classifier performed equally well as functional features. However, the empirical selection of the number of GMM mixtures was a barrier to practically using this method. On the other hand, functional features consistently performed well with and without feature selection methods in both eye and head modalities.

Several feature selection methods were tested using the T-statistic as filtering and using PCA as feature transformation. Using mutual ETF features that are being selected based on the intersected features of all cross-validation turns performed the highest in all classification tasks for both modalities. Moreover, using a variable 98%

of PCA in each cross-validation turn performed slightly better than using a fixed number of PCs of at least 98% of PCA in all classification tasks for both modalities, which indicates better feature representation in each cross-validation turn.

The gender of the subject analysed has a noticeable influence on recognising depression from eye movements and head pose. Females have a higher depression recognition rate than males, especially with the eye modality. With the head modality, the depression recognition rate was statistically similar in both gender groups (only slightly higher in females). The higher recognition rate in depressed females is in line with previous gender differences studies. Such differences in the eye modality may indicate behavioural differences between genders. However, the similarity in the head modality could indicate a physical abnormality such as psychomotor retardation associated with depression.

Moreover, the investigation of expressing positive emotions in depressed and control subjects resulted in remarkable differences between both groups from both eye and head modalities. The recognition rate of expressing negative emotions was statistically lower than expressing positive emotions, which indicates that depressed and controls express negative emotions in a similar manner. The high recognition of depression using positive emotions could conclude that positive emotions are expressed less often in depressed subjects, and that negative emotions have less discriminatory power than positive emotions in detecting depression.

However, even with the reduction of the sample size in both gender-dependent and the expression-dependent experiments, eye and head modality features gave relatively good depression recognition rates (up to 90% AR for gender-dependent and up to 80% AR for expression-dependent), which could be explained by the thin-slicing theory.

Based on these findings, I conclude that eye activity and head movements in general are significantly different between depressed and healthy subjects, and could be used as complementary features for detecting depression. I assume that when fusing these two modalities with the speech modality from the previous chapter, it will not only increase the recognition results, but also increase the confidence in the recognition result. Therefore, the next chapter investigates the influence of fusing speech, eye, and head modalities on depression recognition, and investigates the contribution of each modality to the final results.

Chapter 6

Multimodal Fusion

The previous two chapters investigated unimodal depression detection in several modalities. Chapter 4 investigated speech style and speech prosody modalities individually, while Chapter 5 investigated eye activity and head pose modalities individually. Moreover, several feature selection techniques have been investigated with each modality, where mutual ETF feature selection performed the highest compared to other feature selection methods in each modality. Mutual ETF are the features that exceed a statistical threshold set in advance by a t-value corresponding to an uncorrected p-value of 0.05 ($p < 0.05$), and that intersect in all leave-one-out cross-validation turns.

While individual modalities analysed in the previous two chapters gave reasonable classification results to detect depression, fusing these modalities might not only increase the classification results, but also increase the confidence level in the classifier decisions. This chapter addresses the research question Q3 as stated in Chapter 1. In this chapter, fusion techniques are investigated such as: feature fusion, score fusion, decision fusion, and hybrid fusion, which are detailed in Section 6.1. The results of each fusion technique are presented and discussed in Section 6.2. Moreover, I attempt to explore the reasons behind the classification errors by linking the classification results with the meta-data of subjects in Section 6.3. Finally, different classifiers are compared and fused to examine the robustness of the selected features and to examine the generalisation ability across classifiers in Section 6.4.

6.1 Fusion Methods

As mentioned, multimodal fusion of different modalities can improve the classification result, as it provides more useful information compared to that obtained from a single modality.

Previous chapters elaborated on feature preparation and extraction, which are summarised in Figure 6.1. In Chapter 4, two verbal modalities have been investigated: speech style and speech prosody, where manual labelling and a voice activity detector were used for preparation to extract those modalities' features. In Chapter 5, two nonverbal modalities have been investigated: eye activity and head pose, where subject-specific active appearance models for the eye and face, respectively, were

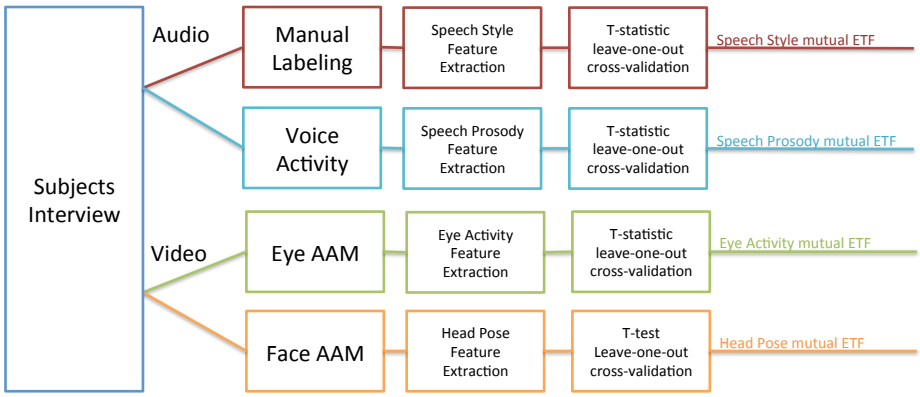


Figure 6.1: Summary of the feature preparation, extraction, and selection

used to prepare for feature extraction. All features from all four modalities undergo several feature selection methods and it was found that mutual ETF performed the highest compared to other feature selection methods. Mutual ETF are the features that exceeded the t-statistic and intersect in all leave-one-out cross-validation turns. Therefore, in this chapter, only classification and fusion of mutual ETF are used for fusion methods investigation.

As discussed in Section 2.2.7, fusion techniques can be performed as prematching (early) fusion and postmatching (late) fusion, as well as a hybrid of both. Since one of the main objectives of this study is to investigate the best fusion approach for the classification of depression, several levels of fusion are experimented on as follows:

1- Early Fusion: is executed either by concatenating the low-level features (data-level fusion), or by concatenating the extracted functional features (feature fusion). Only feature fusion is experimented on in this chapter, due to several reasons: (1) the fact that the sampling rate of the audio is different from the sampling rate of the video, (2) the number of features in each frame (dimensionality) for the audio and video channels differ, (3) all the modalities performed better using the functional features than the low-level ones, and finally (4) that no low-level features exist for speech style features. The differences in sampling rate and dimensionality between audio and video could make the data fusion imbalanced, which would introduce a bias in the classifier towards the larger modality.

a- Feature-level Fusion: where the extracted features of different modalities are combined before classification. Since both verbal and nonverbal modalities are synchronised, this type of fusion is appropriate. Moreover, the extracted functional features from each modality have a temporal nature to overcome normalisation issues, which I believe makes using this fusion method suitable in this case. One of the drawbacks of this fusion method is the increase of feature-vector dimensionality, which might lead

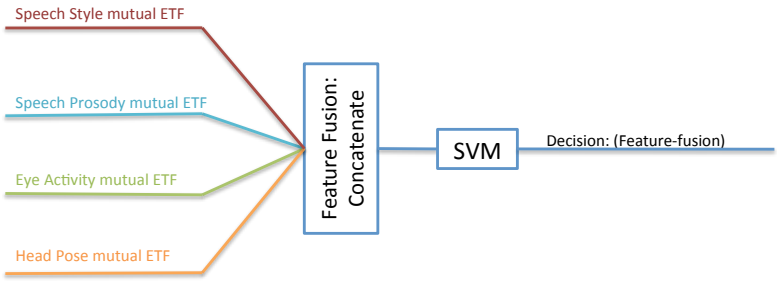


Figure 6.2: Feature fusion method

to a biased decision towards the larger feature vector. That is, if a verbal modality, for example, had substantially more features than the features of a nonverbal modality, the classifier decision is potentially bias towards the verbal modality. However, feature selection methods are used to reduce feature-vector dimensionality and to remove irrelevant features. Furthermore, Min-Max normalisation is used to ensure that all features are unified and to reduce classifier bias towards the higher value features. Since mutual ETF performed the highest compared to other feature selection methods in the previous chapters, in this work, for the feature fusion method, mutual ETF from each modality are concatenated. Figure 6.2 illustrates the feature fusion method used in this investigation.

2- Late Fusion: where the fusion is performed after the classification of each individual modality. Late fusion is performed using either the classifier output scores (score fusion) or labels (decision fusion). Both methods are investigated here.

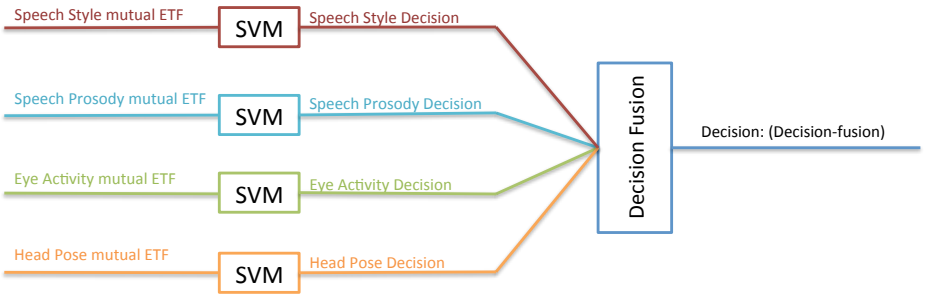


Figure 6.3: Decision fusion method

a- Decision-level Fusion: decisions (labels) from each modality classification are fused. The fusion can be performed either by using logical operators (e.g. AND, OR), majority voting or a secondary classifier. In this work, decision fusion is performed using majority voting, logical AND, and logical

OR, as well as a secondary SVM classifier (see Figure 6.3 for illustration of decision fusion method).

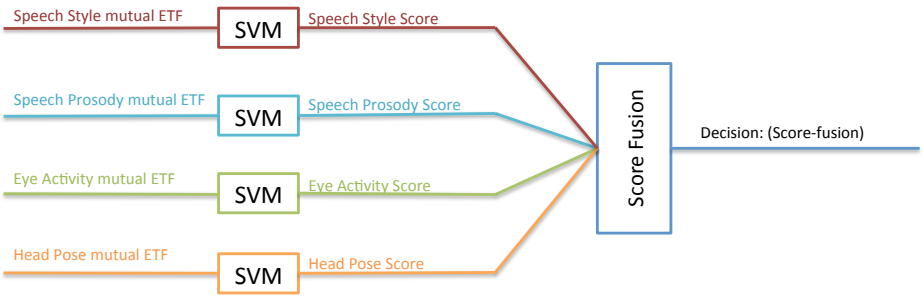


Figure 6.4: Score fusion method

b- Score-level Fusion: where scores from each modality classifier are fused. Since the same classifier is used for each modality, fusing scores from different modalities is simple and no further score normalisation is required. Each modality outputs a confidence score, which could be combined using: mathematical operations (e.g. weighted sum or weighted product), or a secondary classifier. Several mathematical operations are implemented in this work for score fusion, which are the sum-rule, product-rule, and max-rule, as well as a secondary classifier using SVM. The scores in the latter case are the distance from the SVM hyper-plane (see Figure 6.4 for an illustration of the score fusion method).

3- Hybrid Fusion: A hybrid fusion can employ the advantages of both early and late fusion strategies by using the correlation and synchronisation between modalities. In hybrid fusion, the classifier decision from the concatenated features (feature fusion) is fused with classifier decisions from individual modalities. In this work, the decisions from feature fusion and individual modalities are fused using majority voting and a secondary SVM classifier in one-level and two-level hybrid fusion. Either way, a feature fusion of all modalities is performed first to create a new fused modality, which is then treated as an individual modality.

a- One-level: In the one-level hybrid decision fusion, decisions of individual modalities as well as the new fused modality are fused using one level of the decision fusion method (see Figure 6.5, bottom).

b- Two-level: In the two-level hybrid decision fusion, decisions of individual modalities are fused first in a first level, then the resulting decision is fused with the decision of the new fused modality using a second level of decision fusion method (see Figure 6.5, top).

4- Classifier Fusion: employs different classifiers for the same modality, then fuses the scores or the decisions from these classifiers using score or decision fusion

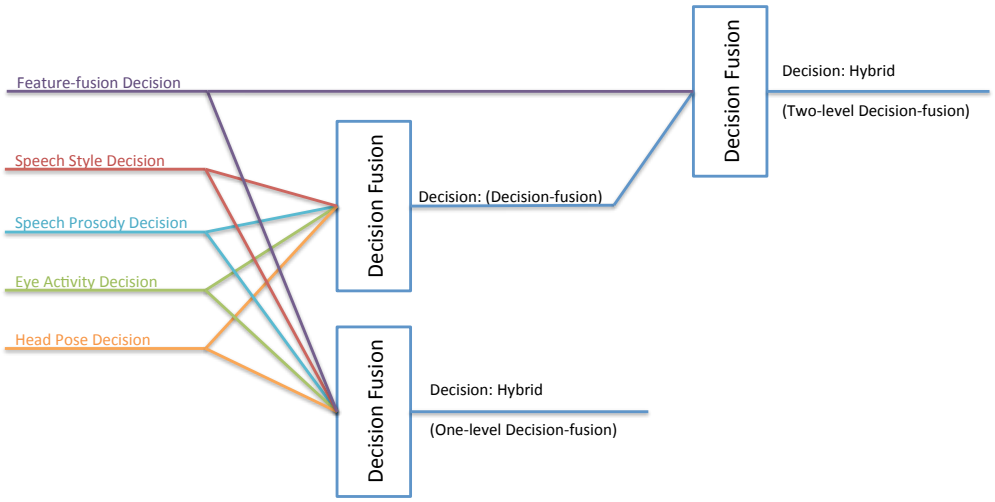


Figure 6.5: Hybrid fusion approaches (one-level and two-level)

methods. Section 6.4 elaborates on the selected classifiers, their comparison and fusion results.

Since larger data is not available to us for this task, the weighted and complex fusion approaches could not be implemented in this study. Future work will include such fusion approaches, as the data recording at the BlackDog is ongoing.

6.2 Fusion Results

Fusion approaches used in this work differ in when and how the modalities in question are fused. Table 6.1 recites the classification results from individual modalities using mutual ETF feature selection (just as a reference for comparison), as well as shows the fusion results using different fusion techniques.

The feature fusion method (see Figure 6.2) results in a slight improvement (+1.7% absolute AR) from the highest performance of individual modalities (speech style and eye activity modalities). Looking at the depression recall, feature fusion has no improvement from the highest individual modalities. However, control recall improved slightly from individual modalities. Even though feature fusion only results in a slight improvement from individual modalities, it it supports and confirms the classification results from both individual modalities and the fused one.

Where the distances from the SVM hyper plane are used as scores, score fusion is explored using the sum-rule, product-rule, max-rule and secondary SVM methods. While product-rule and max-rule score fusion methods have a catastrophic result (lower result than the lowest result of individual modalities), the sum-rule and secondary SVM score fusion methods show a slight improvement (+3.3% absolute) from the highest individual modalities result. Similarly to feature fusion, the sum-rule score fusion method slightly improved the control recall result, while secondary

Fusion Type and Method		Number of Features	Depressed Recall	Control Recall	Average Recall
Individual Modalities	Speech style	41	83.3	86.7	85.0
	Speech prosody	45	80.0	86.7	83.4
	Eye activity	13	83.3	86.7	85.0
	Head pose	7	80.0	70.0	75.0
Feature Fusion	Concatenate	106	83.3	90.0	86.7
Score Fusion	Sum-rule	4 scores	83.3	93.3	88.3
	Product-rule		60.0	43.3	51.7
	Max-rule		56.7	90.0	73.3
	SVM		86.7	90.0	88.3
Decision Fusion	Majority voting	4 votes	83.3	96.7	90.0
	Logic AND		46.7	100.0	73.3
	Logic OR		100.0	46.7	73.3
	SVM		90.0	90.0	90.0
Hybrid: one-level	Majority voting	5 votes	93.3	90.0	**91.7**
	Sum-rule	5 scores	80.0	90.0	85.0
	SVM	5 votes	93.3	90.0	**91.7**
Hybrid: two-level	Majority voting	4-2 votes	73.3	96.7	85.0
	Sum-rule	4-2 scores	80.0	90.0	85.0
	SVM	4-2 votes	93.3	90.0	**91.7**

Table 6.1: Classification results for fused modalities using different fusion methods

SVM score fusion method slightly improved both depressed and control recall results.

Worth noting is that signs (positive and negative scores) are used to identify the class that the subject is classified as belonging to along with the score, which is the distances from the SVM hyper plane. That is, a positive score is given to the depressed class and a negative score is given to the control class. Therefore, mathematical operations that rely on the sign of the scores (i.e. max-rule and product-rule) of individual modalities affect the mathematical sign of the fused modality. For example, if a control subject is misclassified as depressed even for only one modality (a positive sign), the max-rule fusion classification will result in classifying that subject as depressed regardless of the classification of the other modalities. The same applies for the product-rule fusion, where a negative sign in the multiplication operation would have an influence on the final result. Therefore, the catastrophic results obtained by the product-rule and max-rule score fusion methods might be due to the affect of these mathematical operations on the acquired signs of the scores.

On the other hand, decisions out of individual modality classifications are also fused using majority voting, logic AND, logic OR, as well as a secondary SVM. While logic AND and logic OR decision fusion methods had a catastrophic result, majority voting and secondary SVM decision fusion methods had a remarkable improvement (+5% absolute) from the highest individual modalities result. Similar to feature fusion and sum-rule score fusion methods, the majority voting decision fusion method improved on the control recall result. Also, comparable to secondary SVM score fusion, the secondary SVM decision fusion method slightly improved on both depressed and control recall results. The catastrophic results obtained by logic AND and logic OR decision fusion methods are due to the fact that only a few subjects (46.7% recall) have a decision agreement from all individual modalities, hence the

100% recall of either depressed in logic OR, or controls in logic AND. The 100% control recall for logic AND shows that none of the control subject has full agreement from all individual modalities to be a depressed subject. The same applies for the 100% depressed recall for logic OR.

Hybrid fusion, combining both early and late fusing, is examined using decision, score and secondary SVM in two different ways/levels. First, using one level of decision, deals with the feature fusion classification results as a modality along with individual modalities for fusion. Second, using two levels of fusion, the feature fusion classification results are fused with the combined results from a first-level of decision from individual modalities. See Figure 6.5 for a visual illustration. Majority voting and a secondary SVM for decision fusion, and the sum-rule for score fusion are used for both one-level and two-level hybrid fusion since they performed best in the previous score and decision fusion (see above).

Comparing one-level and two-level hybrid fusion with individual modalities classification results (see Table 6.1), both hybrid fusion methods performed either similar or considerably higher than individual modalities' results. The performance of majority voting for one-level hybrid fusion gave a remarkably higher result than individual modalities as well as the feature fusion result, where the depression recall noticeably increased. This high result might be caused by having more votes, which increase the confidence level, and having an odd number of votes to decide upon, which conforms the final vote.

Knowing the risk of overfitting, one- and two-level hybrid fusion were performed using secondary SVM on decisions from individual modalities and the feature-fused modality. Moreover, secondary SVM for both one- and two-level hybrid fusions have considerably higher results compared to individual modalities. On the other hand, the sum-rule of one- and two-level hybrid fusion has similar average recalls to individual modalities, with a slight decrease of depressed and a slight increase of control recalls, respectively.

On the other hand, majority voting of two-level hybrid fusion resulted in a remarkable decrease in depressed recall and an increase of control recalls from individual modalities, resulting in imbalanced recalls, which might be caused by the second level of decision deciding on two votes, which makes it similar to a logic AND.

Comparing the one- and two-level of hybrid fusion, in general, one-level hybrid fusion performed equal to two-level hybrid fusion, with the exception of majority voting decision fusion, where one-level hybrid fusion performed considerably higher than the two-level one. That might be due to the fact that the two-level majority voting decides between two votes: the feature fusion vote and the decision fusion of individual modalities votes, which makes it similar to the logic AND.

Although the result of majority voting for one-level hybrid fusion is equal to the results obtained by secondary SVM from both one- and two-level hybrid fusion, majority voting might be more robust to overfitting. Therefore, the next two sections, concerning classifier errors analysis and classifiers comparison, will report on the majority voting of one-level hybrid fusion for the fused results. Moreover, as the

next chapter deals with the generalisation ability of the selected methods, one-level majority voting of hybrid fusion will be used as the fusion method.

6.3 Classifier Error Analysis

For a better understanding of the classifier misclassifications with each individual modality and the fused modalities, classifier errors are analysed based on subjects' meta-data (e.g. clinical diagnosis, age, etc.) and recordings (e.g. quality noise, illumination, etc.). Table 6.2 shows the number of subjects that have been misclassified in each modality.

Group/Modality		Speech Style	Speech Prosody	Eye Activities	Head Pose	Feature-Fusion	Hybrid: one-level
Depressed	Males	4	6	3	4	5	2
	Femals	1	0	2	2	0	0
Control	Males	2	2	3	5	2	2
	Femals	2	2	1	4	1	1

* The numbers shown are out of 15 subjects per gender per class.

Table 6.2: Number of misclassified subjects in each modality

As can be seen, for each modality, the chance of misclassifying males is higher than for females for both depressed and control groups. Worth noting is that number of subjects who have been correctly classified in all modalities is 28 subjects, half of whom are depressed. Moreover, only 3 of the depressed subjects who are correctly classified in all modalities are males. This result confirms previous conclusions of gender differences [Nolen-Hoeksema, 1987] that depression in women may be more likely to be detected than in men. This might be related to the fact that women are more likely to amplify their mood [Nolen-Hoeksema, 1987]. The same study suggested that men are more likely to engage in distracting behaviours that dampen their mood when depressed, however, that does not explain the misclassification of male control subjects.

For the speech style modality, as the features are behavioural in nature (e.g. response time, pauses, etc.), signal quality and gender-dependent feature issues are eliminated. However, the number of misclassified men in both groups is twice the number of misclassified females. On the other hand, speech prosody features have been normalised to reduce recording and gender differences (as described in Section 4.3), and yet the number of misclassified men is four times higher than for female. Worth noticing is that, none of the depressed women has been misclassified using speech prosody features. Even though the number of misclassified control subjects from speech prosody is equal to the number of misclassified control subjects using speech style, these subjects are not the same (see Appendix A for details).

Moreover, extensive manual labelling of the recordings eliminated background noise and overlapped speech, as well as laughs, coughs, pauses, etc. Also, since differences from recordings and gender are reduced with normalisation, I believe that the misclassifications from speech prosody and speech style are not caused by

the recording quality or the used methods.

With the eye and head modalities, the effect of video quality and whether the subject is wearing glasses were explored. Regarding video quality, only three videos had slightly blurred images, possibly caused by incorrect camera focus (all from the control subset). All three videos have misclassifications from the head modality and only one has misclassifications from the eye modality. However, that does not explain the misclassifications from normal quality videos. Besides, as the eye AAM and face AAM were annotated and trained in a subject-dependent manner, they are also dependent on the recording conditions, which reduces the effect of recording environment and quality differences. Therefore, I believe that the misclassifications were not due to the quality of the videos or the method used.

Looking at the feature fusion modality misclassifications, the feature fusion reduces the classification errors compared to individual modalities, with the exception of depressed men (see Appendix A for details). This result might indicate that the correlation between modalities' features helped to overcome classification errors of individual modalities. Worth noting is that most subjects, who have been correctly classified in two or more individual modalities, have also been correctly classified with feature fusion with exceptions. The first exception is that three depressed men, who have been misclassified with feature fusion have been correctly classified in three individual modalities. The second exception is that three subjects (one depressed man and two controls (a man and a woman), who have been misclassified with feature fusion, have been correctly classified in two individual modalities. Noting that the combination of the two modalities' misclassifications for these three subjects is random. No obvious conclusion could be suggested.

Hybrid fusion overcomes classification errors from individual and feature fusion modalities, where at least three modalities have to agree on a classification decision. As can be seen from Table 6.2, a total of 5 subjects have been misclassified in three modalities or more. Two of these subjects have been correctly classified from only one modality, while the rest have been correctly classified from two modalities (see Appendix A for details).

Errors based on age, diagnosis, depression score, medications (current and history), family history, smoking and alcohol consumption were looked at, none of which had an effect on the classification errors. Moreover, Australia is a multicultural country; therefore, even with selecting native Australian English speakers, three subjects have an appearance of Asian descent (all control subjects: 1 older male and 2 young females). While none of the Asian young females were misclassified in most modalities (only one of the females was misclassified using the head modality), the Asian male was only correctly classified using the speech style modality. As there is not enough data to draw a conclusion, future work could investigate the influence of cultural backgrounds.

Therefore, I believe that, as all extracted features are behavioural in nature and normalised, some subjects with certain personalities or cultural backgrounds might act and behave differently from the general characteristic of depression regardless of their mental health. For example, depressed patients, who have more head move-

ment as they speak, are misclassified as control subjects and vice versa. The same applies for the eye and speech modalities. As the current data collection did not include a personality assessment, I could not derive a solid conclusion on this point. Adding a personality assessment is being considered for the ongoing data collection at the Black Dog institute. Nevertheless, I strongly believe that the proposed method works within reasonable accuracy in general behavioural patterns for depressed compared with healthy control subjects.

6.4 Classifier Comparison and Fusion

Automatic emotion recognition approaches have used a variety of classifiers, both descriptive (generative) and discriminative approaches, but it is not clear, which one performs best for the detection of depression. In order to accurately detect depression and identify which classifier performs better for this task, a comparison of two further classifiers (beside the previously investigated SVM) is performed: (1) multilayer perceptron neural networks are a popular classifier from the literature, and (2) the relatively new Hierarchical Fuzzy Signature classifier (see Section 2.2.6). In this section, besides comparing the classifier performances, classifier fusion is also investigated. Classifier fusion is performed in two methods to investigate which performs better for this task, namely fusion at the modalities level or at the classifiers level.

6.4.1 Classifier Comparison

In this section, similar to previous chapters, all three classifiers are employed in a binary (i.e. depressed/non-depressed) subject-independent scenario. Moreover, to mitigate the effect of the limited amount of data, a leave-one-subject-out cross-validation was used in all the classifiers without any overlap between training and testing data. For a fair comparison, mutual ETF features from each modality are selected and used. Furthermore, I acknowledge that the selected features, which are the mutual ETF, were selected based on the comparison of another five methods of feature selection using an SVM classifier. Therefore, the selected features are optimized on the SVM, which might have an effect on the classification results of the other classifiers. However, for consistency and comparison with the previous investigations, as well as focusing on the research questions, I chose to use the mutual ETF regardless of the potential effect on the other classifiers results. Investigating the effect in more detail is an interesting task to be explored in the future.

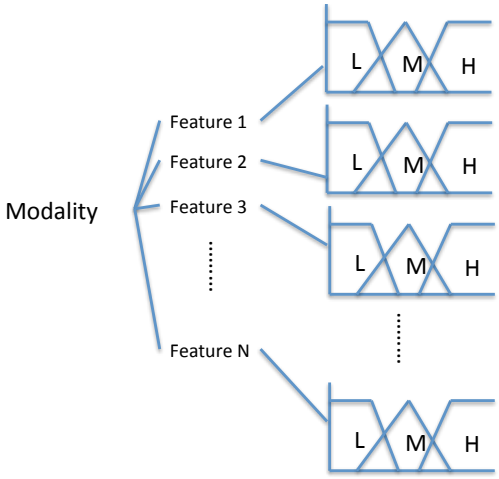
Table 6.3 shows the classification results from individual and feature fusion modalities as well as the hybrid fusion result for the three classifiers. The SVM results are recited as reference for easy comparison with the classification results using HFS and MLP classifiers.

A brief description of HFS has been given in Section 2.2.6. I hypothesise that, given the continuous range nature of emotions in general and the potential overlap between them, fuzzy systems might be suitable for the task. However, the choice of the membership function, the aggregation function, and the number of fuzzy sets are

Modality/Classifier	SVM	HFS	MLP
Speech style	85.0	83.3	81.7
Speech prosody	83.3	65.0	70.0
Eye activity	85.0	76.7	71.7
Head pose	75.0	55.0	60.0
Feature-fusion	86.7	58.3	80.0
Hybrid: one-level	91.7	83.3	83.3

Table 6.3: Correct classification results (in %) when using different classifiers

critical for getting accurate results. In this section, the HFS construction approach is adopted based on the Levenberg-Marquardt method (Mendis et al. [2006]). The fuzzy signature was constructed using the mutual ETF as the branches. For each modality, each feature (branch) represents a fuzzy value calculated using Fuzzy C Mean (FCM) clustering into three fuzzy sets (Low - Med - High) (see Figure 6.6). That is, in each leave-one-out turn, the training set is used to construct three fuzzy sets for each feature of each modality. Based on these fuzzy sets, training membership values are calculated for each fuzzy set for the training features. Then, the testing set is fitted into these three fuzzy sets of each feature to output testing membership values as well. The training membership values are used to create the fuzzy model, and then the testing membership values are tested against the created model to output a testing label, which is compared to the actual label to calculate the accuracy.



L, M, H: Low - Med - High fuzzy sets, respectively

Figure 6.6: Constructing fuzzy signature

The modalities and fusion classification result using HFS classifier shown in Table 6.3 implies that the HFS classifier is a reasonable choice for depression detection. All individual and fused modalities gave above chance level accuracy; however, head pose and feature fusion modalities gave the lowest results. Even though the hybrid fusion did not improve the result from individual modalities, it was not lower than the highest result. The speech style modality gave a high result compared to other

modality, which is similar to the result when using SVM as classifier. This finding implies that speech style features hold discriminative features to distinguish depression.

The MLP has been briefly described in Section 2.2.6. An MLP using two hidden layers was implemented. The first layer contains half the number of features in the modality as perceptrons, and the second layer contains one sixth ($\frac{1}{6}$) of the number of features in the modality as perceptrons. The number of perceptrons was chosen empirically. The input for the MLP was the mutual ETF and the target output was the binary label of the classes (1 for Depressed, 0 for Control). The parameters used to create the MLP in this work are: Levenberg-Marquardt as the training function, hyperbolic tangent sigmoid as activation function for hidden layers, mean squared error as a cost function. To reduce the effect of the random elements of MLP training on the final results, an ensemble of 100 MLP was implemented, where the final result is based on the majority voting of these 100 networks.

As shown in Table 6.3, MLP shows reasonable classification results with all individual and fused modalities giving an above chance level accuracy. Similar to using SVM and HFS, the speech style modality performed the highest result compared to other individual modalities, which supports the previous finding that speech style contains strong distinguishing features to detect depression. The hybrid fusion result of MLP slightly improved from individual modalities.

Comparing the three classifiers, SVM performed better than HFS and MLP classifiers in every individual modality and fused modalities. This finding indicates that using SVM is suitable for the extracted features and methods used for the application of depression detection.

6.4.2 Classifier Fusion

Classifier fusion was also investigated to show its effect on the overall task of depression detection. While classifier fusion is usually performed as late fusion, either by decision fusion or score fusion, scores of different classifiers might not be compatible with each other and further score normalisation would be needed before fusion. Therefore, the classifier fusion was performed using majority voting of classifiers' decisions. Since several modalities are investigated, as well as classifiers, the fusion can be done in two methods: (1) at classifiers level, and (2) at modalities level as discussed in following.

The first method is fusing decisions at the classifiers level. That is, after fusing decisions from different modalities of the same classifier, the final classifier decisions are fused, as illustrated in Figure 6.7. This figure, also shows the accuracy classification results from each modality and from each fusion level.

As seen in Figure 6.7, the first level of fusion are the hybrid fusions for the classifiers, which have been discussed earlier in classifiers comparison in the previous section. The second level of fusion is the final classifier fusion result, which performed 88% average recall. This classifier fusion result is higher than individual HFS and MLP classifiers results, but less than the SVM classifier result. A majority agree-

Figure 6.7: Classifier fusion in the classifiers level approach *(percentages between parentheses are the accuracy classification results in terms of AR from each modality and from each fusion level)*

ment between classifiers defines the final classification result. The fact that the final result is based on the three classifiers, implies that the misclassifications in the three classifiers do not necessarily agree with each other, especially the misclassifications from HFS and MLP. This finding also supports the previous finding, that SVM is a suitable and reliable choice for the application of depression detection.

The second method is fusing decisions at the modalities level. That is, different classifier decisions are fused for each modality, then the modality decisions are fused to get the final decision. Figure 6.8 illustrates the second method and shows the accuracy of the classification (AR) results at each fusion level.

As shown in Figure 6.8, the first fusion level fuses the decisions from the three classifiers for each modality. For each modality fusion, the fusion results substantially improve from individual HFS and MLP classification results and either keep or slightly decrease the classification results of SVM. This finding supports the previous finding of the fusion at classifiers level, where there is less agreement in misclassifications between HFS and MLP compared with their agreement in SVM misclassifications. The second level of classifier fusion gives the final result, which resulted in 87% average recall. This performance is equal to the higher performance of modal-

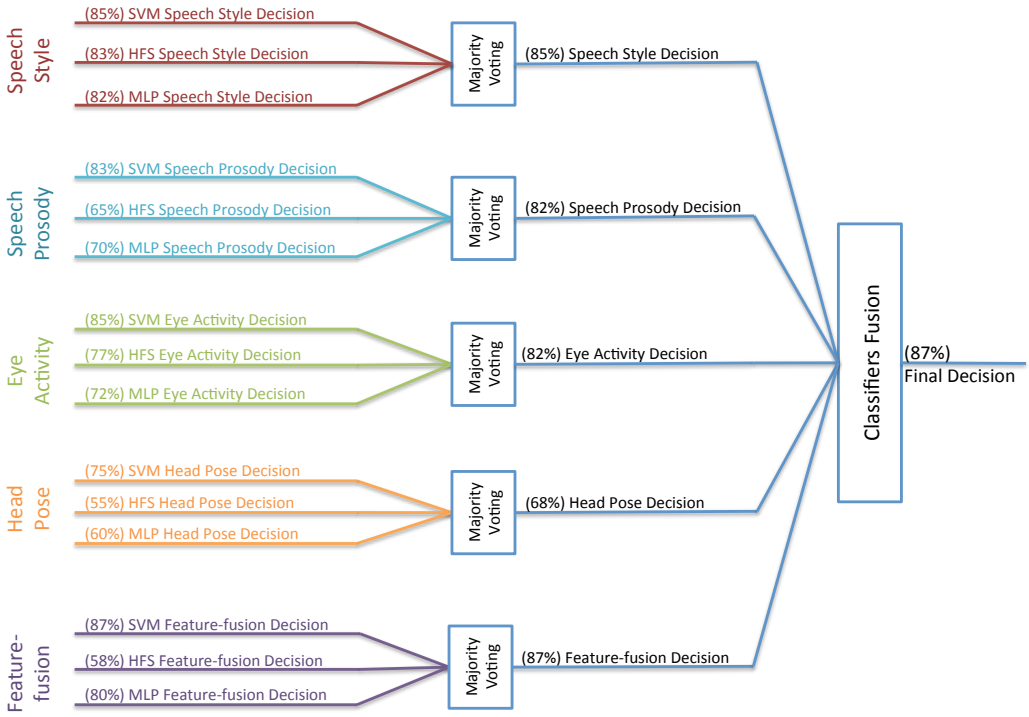


Figure 6.8: Classifier fusion in the modalities level approach *(percentages between parentheses are the accuracy classification results in terms of AR from each modality and from each fusion level)*

ities fusion, which was performed by feature fusion, although the depression and control recalls differ.

The fusion results from the two methods of classifier fusion (at classifiers level and at modalities level) are similar, with the modalities level being slightly lower. However, the depression and control recalls differ between the two classifier fusion methods. Even with having the same inputs, using different fusion combinations and levels affects the final results, which implies inconsistency in decisions between the three classifiers. The inconsistency issue needs more investigation of classifiers' modeling and parameter choices. Therefore, SVM seems to be a more suitable choice for the selected methods and the task of detecting depression.

6.5 Summary

Intending to ultimately develop an objective multimodal system that supports clinicians during the diagnosis and monitoring of clinical depression, fusing verbal and nonverbal temporal patterns of depression was investigated. Verbal modalities (speech style and speech prosody) and nonverbal modalities (eye activity and head pose) were investigated individually in previous chapters. The results from individual modalities were encouraging for the automatic classification and detection

of depression on their own. Therefore, the performance of the classification was examined when fusing these modalities, hypothesising an improvement when fused. Knowing that fusion could be performed at different stages using different methods, several fusion methods were explored for comparison and determining the best method for the task of detecting depression.

Several methods of feature, score, decision, and hybrid fusions were investigated using SVM classifier. Simple concatenation was used for feature fusion, where a slight improvement from the highest individual modality was acquired. I believe that feature fusion needs features that are compatible in nature, which could be acquired in the feature extraction stage as well as using normalisation methods before concatenation, which was the case in this study. Several score fusion methods were investigated, where the sum-rule and a secondary classifier resulted in a slight improvement from the highest individual modality. The product-rule and max-rule of score fusion had catastrophic results (lower than the lowest individual modality), which might be due to the effect of the mathematical operations on the score sign, which indicates a need for pre-normalisation of the scores. Decision fusion was investigated using several methods, where majority voting and a secondary classifier led to a remarkable improvement from individual modalities.

Moreover, hybrid fusion was performed using one- and two-levels, with the highest fusion results achieved using majority voting and a secondary classifier. While both hybrid fusion methods had an improvement from individual modalities, one-level hybrid fusion had a remarkably higher result from two-level hybrid fusion using the majority voting method. That might be due to the number of votes in the one-level hybrid fusion being a higher and an odd number compared to the two-level one. As a secondary classifier of hybrid fusion might risk overfitting, it would likely need a larger database for this approach to be convincing and valid. I believe that the majority voting one-level hybrid fusion is more reliable and more robust to overfitting than the two-level one and the secondary classifier method. Therefore, majority voting one-level hybrid fusion will be selected as the fusion method for further investigations including generalisation (see next chapter).

The classification errors were analysed for a better understanding of the proposed method for detecting depression. In line with the literature, depression in women was more likely to be correctly classified than in men from both groups (depressed and control). Several technical issues that might have an effect on the classification were eliminated, including audio and video quality, gender-dependent features, and recording conditions by using normalisation techniques and considering the type of extracted features. Moreover, the subjects' meta-data were analysed, but showed no effect on the misclassifications. Therefore, I strongly believe that the proposed method works on general behavioural patterns of depressed subjects compared with healthy control subjects. In future studies and data collections, investigations of the effect of personality and cultural background are needed to further analyse the misclassifications.

As another way of fusion, three classifiers' decisions (SVM, HFS, and MLP) were compared and fused. Such comparison helps to generalise the findings of each

modality and to investigate their robustness across classifiers. While comparing modalities classification results, the speech style modality consistently gave high results across the three classifiers. This finding implies that speech style features are robust and have strong characteristic to detect depression. Comparing the three classifiers, SVM performed better than HFS and MLP, which their performances being equal. Fusing classifiers' decisions was investigated in two ways: at classifiers level and at modalities level. Their performance was similar to each other. The classifier fusion result was slightly lower than using SVM alone, but slightly higher than using HFS or MLP. The agreement between HFS and MLP was inconsistent with each other, but mostly they agreed with the SVM decision. Such inconsistency needs further investigation and might indicate a need for better structures and modeling as well as parameter choices for HFS and MLP classifiers. Therefore, SVM seems to be a more suitable choice for the selected methods and the task of detecting depression.

To this end, I conclude, based on the BlackDog dataset, the best method of fusing the verbal and nonverbal modalities for the task of depression detection investigated here was found to be a hybrid fusion using a one-level majority voting technique. However, such conclusion could not be derived using only one specific dataset. Therefore, the coming chapter will apply all concluded methods previously investigated and used on the BlackDog dataset on two different datasets, which are not only different in specifications such as languages and recording environment, but also different in terms of depression diagnosis methods. Such an investigation is aimed at testing the ability to generalise the proposed methods to different datasets. The results then are expected to provide further insight and possibly support for the conclusions derived from the BlackDog dataset.

Chapter 7

Generalisation Across Datasets

The previous chapters examined automatic depression detection on the BlackDog database using several modalities. Investigated modalities included speech style and speech prosody, as well as eye activities and head pose, where several feature types and feature selections were compared. Furthermore, these modalities have been explored individually and when fused using several fusion techniques. Such investigations resulted in proposing a complete system with a high accuracy in detecting depression using the BlackDog dataset. However, to validate the usability and generalisability of the proposed system (see Section 7.1.1), it should be applied it on other different datasets.

In this chapter, the proposed system is applied to two different depression datasets that are different in language, depression scale and recording environment, in order to measure its generalisability, which addresses Q4 of the research questions of Chapter 1. These datasets are the University of Pittsburgh depression dataset, and the Audio/Visual Emotion Challenge depression dataset. First, the proposed system is described in Section 7.1.1, which consist of best performing methods investigated using the BlackDog dataset. Since both AVEC and Pitt datasets differ from BlackDog dataset in language, depression scale and recording environment, applying the proposed system on these datasets involves some adjustments in modality preparations as described in Sections 7.1.2.

Second, the system is applied on the AVEC dataset and Pitt dataset individually as well as with different dataset combinations. The results are discussed in two parts: results of applying the system on individual datasets in Section 7.2, and when applied on a combination of the three datasets in Section 7.3 (generalisation).

7.1 Proposed System

7.1.1 Description of this proposed system applied on BlackDog dataset (a review)

Using the BlackDog dataset, both individual modalities (Chapters 4 and 5) and fused modalities (Chapter 6) compared several methods to propose the best system to detect depression from verbal and nonverbal cues. In this section, the methods for the proposed system are reviewed and summarised.

Table 7.1 summarises the selected methods of the proposed system performed on the BlackDog dataset. Four modalities were investigated: speech style, speech prosody, eye activities, and head pose, where each modality goes through preparation and normalisation stages before features could be extracted, then processed for detecting depression, as follows:

	Selected Method	Preparation	Pre-Normalisation
Investigated modalities	Speech style	extensive manual labelling - speech rate over subject's segment	rate over speech duration
	Speech prosody	voice activity detector over subject's segment	Z-score
	Eye activity	subject-specific 74-points eye-AAM (trained over 45 annotated images per subject)	rate over distance of other points
	Head pose	subject specific 68-points faceAAM (trained over 30 annotated images per subject)	range of reasonable head movement
Extracted features	functional		
Post-normalisation	min-max		
Feature selection	mutual features that exceeded the T-statistic (refereed to as mutual ETF)		
Classification	binary classification with leave-one-subject-out cross-validation using SVM		
Fusion	one-level hybrid fusion		

Table 7.1: Overview of selected methods of the investigations performed on the BlackDog dataset

Investigated modalities:

Speech style: Speech style features include speakers' turns, response time, pauses, speech rate, etc. (see Section 4.2), and are prepared using manual annotation. Then using a subject's segment, a code to extract speech rate is executed. Since these features depend on the duration of the interview, a pre-normalisation of these features is performed using the duration of entire interview.

Speech prosody: Over a subject's turns, a voice activity detector is executed to extract sounding segments, and then several prosody features are extracted such as F0, MFCC, energy and others (see Section 4.3). As pre-normalisation process, a Z-score normalisation is applied over the extracted low-level feature to reduce any differences such as gender and recording environment.

Eye activities: For preparation to extract eye activity features, which include iris movement and blinks, a subject-specific eye-AAM that contains 74-points around the eye area and the eyebrows was built. Since the BlackDog dataset interview is interactive, many head movements exist, which required annotating an average of 45 images per subject to train the eye-AAM model. Furthermore, to reduce the difference of eye area size and structure, and to reduce the effect of the variability of distance of the cam-

era, a normalisation procedure is performed so that the features extracted are calculated as rate over face structure and angles (see Section 5.1).

Head pose: To extract the three angles of head pose, a subject-specific face-AAM with 68 points was built using 30 annotated images per subject, and then the points were projected on a three dimensions face model. To normalise the angles and eliminate outliers, a specific angle range is restricted within reasonable head movement (see Section 5.2).

Extracted features: In each modality, low-level (frame by frame) features are compared with statistical features extracted from the entire interview, where statistical features performed statistically higher in detecting depression.

Post-normalisation: To reduce classifier bias to high values feature from other features, the features were further normalised using Min-Max normalisation in leave-one-out cross validation manner (see Section 3.1).

Feature selection: Comparing several feature selection methods, it has been found that selecting the mutual features that exceed the T-statistic in each turn of leave-one-out cross validation performed best for the task of detecting depression using each individual modality (see Chapters 4 and 5).

Classification: Classifiers were compared, and it has been found that SVM performed best for this task (see Chapter 6). For classification, leave-one-out cross-validation was used in a binary manner (depressed vs. controls).

Fusion: Fusion techniques were explored and it has been found that a hybrid fusion best fits the task in question, where decisions of feature-fusion of the used modalities and the individual modalities were fused in one-level of majority-voting (see Chapter 6).

7.1.2 Differences of Applying the Concluded System on AVEC and Pitt datasets

As mentioned earlier, the three datasets are different in several aspects (see Table 3.1 for details). These differences could affect the generalisation results. Therefore, attempts to reduce some of these differences have been considered as follows:

- Each dataset uses different depression screening instruments to measure depression severity. To create a common metric of depression, we converted the different metrics to their QIDS-SR equivalents using the conversion table from Depression Scores Conversion [Online], and categorise subjects based on the severity level of depression. For more details about depression severity measure, its mean score and range for each dataset, as well as their QIDS-SR equivalents, see Table 3.1.

- While the BlackDog dataset compares depressed patients with healthy controls, both Pitt and AVEC datasets aim to monitor depression severity. Therefore,

the originally intended classification problem is different for each dataset. To overcome this issue, the subjects in these datasets were categorised into two groups for a binary classification: severe depressed vs. low depressed (AVEC and Pitt)/healthy controls (BlackDog).

- The data collection procedure for each dataset differs. BlackDog uses structured stimuli to elicit affective reactions, which includes an interview of asking specific open ended questions, where the subjects are asked to describe events in their life that had aroused significant emotions, to elicit spontaneous, self-directed speech and related facial expressions, as well as overall body language. The Pitt data collection procedure was conducted by interviews using the HRSD questions, where patients were interviewed and evaluated by clinicians. On the other hand, the AVEC paradigm is a human-computer interaction experiment containing several tasks including telling a story from the subject's own past (i.e. best present ever and sad event in the childhood). Therefore, in this study, only the childhood story telling from AVEC is analysed in order to match the interviews from the BlackDog and Pitt datasets.

- While BlackDog records only one session per subject, AVEC and Pitt have multiple sessions per subject (up to four). In this study, only one session for each subject was selected, thereby splitting the subjects into two groups: severe depressed vs. low depressed/healthy controls. I also aimed to have a balanced number of subjects in each class to reduce classification bias towards the larger classes; hence the relatively small number of selected subjects in each dataset, but this is a common problem in similar studies.

- The duration of segments for each subject in each dataset varied. Therefore, to reduce variability from the length of subjects' segments, temporal features were extracted over the entire segments.

- Recording environment and hardware are also different for each dataset. The audio channel in particular is more vulnerable to the recording environment than the video channel (e.g. microphone distance, background noise, sampling rate, etc.). The video channel has also its obstacles regarding recording environment, such as: lighting condition, cameras (vocal point, type and distance), and video files (resolution, frame rate and dimensions). Therefore, several feature normalisation techniques were applied before and after feature extraction (pre- and post- normalisation), as described in Section 7.1.1.

Since the datasets used in this study differ on several aspects, in particular, recording environment, modality preparations for applying the system would involve a few adjustments. These adjustments are as follows:

Speech style modality: Unlike the BlackDog and Pitt datasets, which include interviews in their data collection procedure, the AVEC dataset recordings are human-computer interactions. There is no interaction with another human in

AVEC dataset. Therefore, some speech style features could not be extracted such as speakers' turns, overlapped speech, response time, etc.

Since the AVEC dataset lacks speaker interactions, where several speech style features could be extracted, and to have equal comparison between the three datasets, the speech style modality will not be investigated in the generalisation study in this chapter.

Speech prosody modality preparations: The AVEC dataset recordings are from a human-computer interaction task and, therefore, no segmentation to separate speakers is required.

On the other hand, the Pitt dataset contains structured interviews with a psychologist and, therefore, speaker segmentation is required before analysing a subject's speech signal. University of Pittsburgh researchers manually transcribed the interviews in the Pitt dataset, where I isolate each subject's segments using the transcript's timeline. Pitt subject segments were extracted for further analysis and feature extraction.

As with the BlackDog dataset, a voice-activity-detector (as described in Section 4.1) is executed on the AVEC and Pitt subject's speech to extract sounding segments as preparation for speech prosody feature extraction.

Eye activity modality preparation: For eye activity feature extraction preparation, the same eye-AAM method used on the BlackDog dataset was used on the AVEC and Pitt datasets. For the AVEC dataset, since the recordings are from a human-computer interaction task, only few head movements exist. Therefore, the required images to annotate for training the eye-AAM were fewer compared with the spontaneous interview in the BlackDog dataset. Only 7 images per subject were annotated on average in the AVEC dataset.

Moreover, since the Pitt dataset contains interviews with a psychologist, many head movements are expected. Therefore, more images need to be annotated to train the eye-AAM with different variations of head poses. For the Pitt dataset, 15 images per subject were annotated on average.

Head pose modality preparation: For head pose feature extraction preparation, instead of using subject-specific face-AAM as used in the BlackDog dataset, on the AVEC and Pitt datasets, generic face fitting for face detection was used. The generic face model uses an optimised strategy of constrained local models [Saragih et al., 2009]. The CLM model used for face detection contains 64-points around the face, where 46 corresponding points were projected to the 58-points of the 3D face model to extract head pose features as described in Section 5.2.1.

The aim of this chapter is to generalise the finding of the proposed system of detecting depression to the BlackDog dataset to other datasets. The generalisation investigation is done by applying the proposed system on the datasets individually as well as in different dataset combinations. The following two sections show the

results of applying the proposed system on the three datasets individually and with different combinations (see Sections 7.2 and 7.3, respectively).

7.2 Results of Generalisation on Individual Datasets

The results of applying the proposed system on the three datasets individually are presented in Table 7.2. Figure 7.1 recites Table 7.2 AR results for visual illustration. For comparisons, the table shows some of the classification results for BlackDog dataset, and presents results of excluding the speech style modality from both feature fusion and hybrid fusion. The table shows the classification results of applying the system on the AVEC and Pitt datasets individually. Worth noting is that Min-Max post-normalisation and feature selection using mutual ETF were applied on each dataset individually as per leave-one-out cross-validation, hence the differences in the number of features between the datasets, as shown in Table 7.2.

Dataset	Modality	Number of Features	Depressed Recall	Control Recall	Average Recall
BlackDog	Speech prosody	45	80.0	86.7	83.3
	Eye activity	13	83.3	86.7	85.0
	Head pose	7	80.0	70.0	75.0
	Feature fusion	65	86.7	90.0	88.3
	Hybrid: one-level	4 votes	83.3	90.0	86.7
AVEC	Speech prosody	62	93.8	100	96.9
	Eye activity	31	75.0	87.5	81.3
	Head pose	29	56.3	75.0	65.6
	Feature fusion	122	100	81.3	90.6
	Hybrid: one-level	4 votes	81.3	93.8	87.5
Pitt	Speech prosody	74	100	63.2	81.6
	Eye activity	20	89.5	94.7	92.1
	Head pose	31	84.2	89.5	86.9
	Feature fusion	125	94.7	79.0	86.9
	Hybrid: one-level	4 votes	94.7	94.7	94.7

Table 7.2: Classification results of individual datasets. The four votes for the hybrid fusion represent the decisions from the three individual modalities and the feature fusion.

Even though the BlackDog dataset classification results were discussed in earlier chapters, a few differences are introduced in this section. As mentioned earlier, the speech style modality was excluded from feature and hybrid fusions, to allow for an equal comparison with the AVEC dataset where there are no human-human interactions to extract speech style features from. Comparing feature fusion *including* the speech style modality (see Table 6.1 in previous chapter) with feature fusion *excluding* the speech style modality (see Table 7.2 in current chapter), the performance of the latter slightly improved (+1.6% absolute). However, comparing hybrid fusion including and excluding the speech style modality, the performance of the latter drops remarkably (-5% absolute) (see Table 6.1 vs. Table 7.2), which might be due to when excluding the speech style modality, the number of votes is an even number, while when including the speech style modality the number of votes for the hybrid

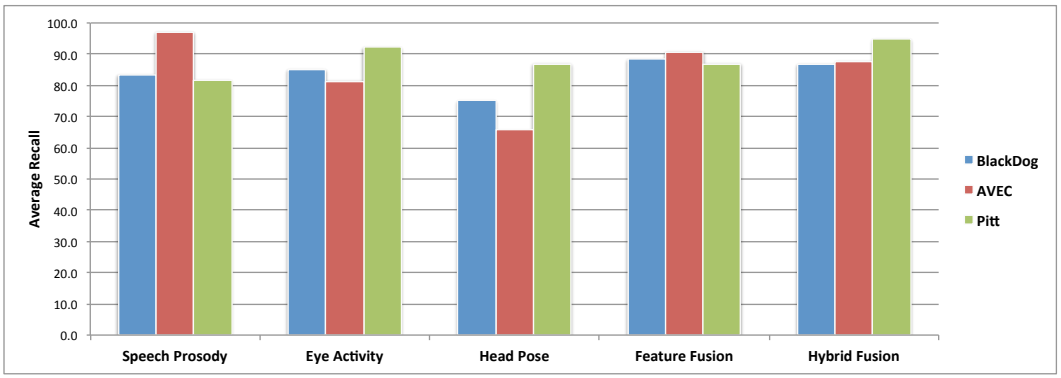


Figure 7.1: Average recall of classification results of individual datasets

fusion is an odd number. Therefore, with an even number of votes, the majority voting treats equal votes as a logic AND, which could have an effect on the results.

All modality classification performances for both AVEC and Pitt dataset are considerably above chance level, performing an average of 82% AR. The classification results from each modality of the three datasets are comparable, which supports the hypothesis that the proposed system has the ability to generalise to different datasets when applied individually on each dataset.

The speech prosody modality (see Figure 7.1), on the other hand, has different classification results for the three datasets. Speech prosody performed best in the AVEC dataset (97%), and equally in the BlackDog and Pitt datasets (82%). The high performance in the classification results in the AVEC dataset might be due to the clear distinction between severely depressed and low depressed patients, as the gap between depression scores for the two groups is very wide (score range of 30-45 for severe depression and score range of 0-3 for low depression using BDI test, see Table 3.1). Such differences in depression severity might have a distinct effect on the patient's vocal cords.

Moreover, eye activity modality classification results perform consistently strongly across all three datasets (see Figure 7.1), which implies that eye activity is a strong characteristic to differentiate severe depressed from low or non-depressed behaviour.

Similar to the BlackDog dataset, the lowest result in the AVEC dataset was obtained by the head pose modality (see Figure 7.1). While BlackDog head pose modalities performed at 75% AR, the AVEC head pose modality performed at 65% AR, which is expected for the AVEC dataset since the recording procedure is a human-computer interaction task, where the head movements are limited. On the other hand, the Pitt head pose modality performed the highest compared to the BlackDog and AVEC datasets.

Feature fusion classification performances are comparable when comparing the feature fusion results of the datasets, but vary when compared with the modalities that it fuses. That is, the BlackDog feature fusion result improved from the modal-

ities it fuses, the AVEC feature fusion result was lower than the highest individual modalities result but higher than the lowest individual modalities result, while the Pitt feature fusion result is equal to the highest individual modalities result. This variation of feature fusion performance is not remarkably different. The reasons behind it could have several sources, such as differences in signal quality, differences in depression diagnosis, differences in data collection procedure, etc. where more data are needed for such investigations. Nevertheless, the classification results of fusing modalities in feature fusion were not catastrophic, that is, not lower than the lowest modality result. That could imply that the fused modalities are correlated and complement each other for the task of detecting depression.

Like feature fusion, hybrid fusion using majority voting of decisions from individual modalities and from feature fusion had a slightly varied outcome compared to the modalities that it fuses for each dataset. For BlackDog, the hybrid fusion results decreased slightly (-1.6% absolute) from the highest individual modality, which in this case was the feature fusion. For AVEC, the decrease in hybrid fusion results was remarkable (-9.4% absolute) from the highest individual modality, which in this case was the speech prosody modality. On the other hand, Pitt hybrid fusion increased slightly (+2.6% absolute) from the highest individual modality, which in this case was the eye activity modality. This variation could imply varied decisions for each subject in the investigated modalities. That is, modalities had varied agreement/disagreement for the same subject. However, as with feature fusion results, the classification results of the hybrid fusion were not catastrophic, which could give a stronger confidence of the hybrid fusion compared to individual modality results.

Depression recalls and low or non-depression recalls were calculated separately (see Table 7.2), not only to measure the average recalls, but also to investigate which group was correctly classified compared to the other. For both the BlackDog and AVEC datasets, the non- and low depression subjects have higher recalls than severe depression in most cases, except for the head pose modality for BlackDog and feature fusion for AVEC, while for the Pitt dataset the severe depression recalls were higher than low depression recalls in all cases. That might be due to the clear difference in non-depression in the BlackDog dataset and very low depression in AVEC compared to the severe depression group, while the severity of low depressed patients selected on the Pitt dataset was higher in severity range from those selected from the low depressed group in the AVEC dataset and the non-depressed group in the BlackDog dataset. Therefore, the differences between low depressed and severe depressed in the Pitt dataset are not clearly distinct, which might be the reason behind the reduction in the classification recalls for low depressed group.

7.3 Results of Generalisation on Combinations of Datasets

Following the success of applying the proposed system on the datasets individually, where the final classification results from all the datasets were comparable and considerably above chance level (average of 86% AR), I attempt to apply the proposed

system on combinations of the three datasets. I acknowledge that the datasets differ in several aspects, most importantly the differences in recording environment, tasks, and in depression severity scales, which could have a large effect of the classification results when combining different datasets. However, I hypothesise that combining the three datasets is beneficial for generalising a model that could detect depression. Moreover, selecting features that could detect depression on the three datasets could be a starting point of having a generalised system regardless of the differences of recording environment and severity scale of depression. Such an investigation could give an insight to the effectiveness and weakness of generalising to different datasets in general and the proposed system in particular.

In this section, selecting the training and testing sets for classification for applying the proposed system on the datasets combinations is done in two ways:

Leave-one-subject-out cross-validation method: To mitigate the limitations of the relatively small amount of data, and also to train the classifiers on varied observations, a leave-one-subject-out cross-validation is used on the combinations of the datasets (as performed with the individual datasets) without any overlap between training and testing data. This method could overcome overfitting the model on the training set, especially as the final selected SVM parameters generalise to all training observations with the leave-one-out cross-validation (see Section 3.1). In other words, the common parameters that give the highest average training accuracy of all training sets in the cross-validation models are picked, hence the need for a wide range search. I believe that this method of selecting the parameters reduces overfitting issues on the training set and therefore assists in generalising to different observations in each leave-one-out cross-validation turn.

Separate train-test dataset method: In this method, one or two datasets are used for training and then the remaining dataset(s) are used for testing. In this method, the SVM parameters are selected based on the highest accuracy of the training set. This method could suffer from overfitting to the training set, and might not generalise to the testing set(s), especially as the testing set is completely different. This method is applied to investigate the generalisation ability of the depression detection method to unseen data.

I hypothesise that when using different combinations of datasets, leave-one-out cross-validation would give a higher performance than the train-test method, because the model in leave-one-out cross-validation is trained over varied samples of combined datasets, which reduces model overfitting to the training set. However, both methods are investigated here to shed more light onto the area of cross-corpora generalisation.

Feature Selection and Normalisation

Since the classification is done in a binary manner, a simple T-test is sufficient for finding significant differences between the two analysed groups and could be utilised

as a feature selection method (see Section 3.1). For the generalisation study, features that exceed the T-statistic are selected for the classification problem in two approaches.

Modality			Feature	BlackDog	AVEC	Pitt	BlackDog+AVEC	BlackDog + Pitt	AVEC + Pitt	All 3 datasets
Speech prosody	HNR	1	range			✓		✓	✓	✓
	Jitter	2	range			✓		✓	✓	✓
	Voice quality	3	variance		✓		✓	✓	✓	✓
		4	delta average	✓	✓		✓	✓	✓	✓
		5	delta minimum	✓			✓	✓	✓	✓
		6	delta range		✓		✓	✓	✓	✓
		7	delta variance		✓		✓	✓	✓	✓
	Log energy	8	range	✓	✓		✓			✓
	Shimmer	9	range		✓		✓		✓	✓
		10	delta maximum	✓		✓		✓		✓
	Formants	11	4th formant range		✓	✓	✓	✓	✓	✓
		12	4th formant variance	✓			✓	✓		✓
		13	5th formant standard deviation	✓			✓	✓	✓	✓
		14	6th formant standard deviation	✓			✓	✓		✓
		15	1st formant delta average	✓		✓		✓		✓
		16	1st formant delta maximum	✓	✓		✓	✓	✓	✓
		17	1st formant delta standard deviation	✓	✓	✓	✓	✓	✓	✓
		18	2nd formant delta standard deviation	✓			✓	✓		✓
		19	4th formant delta range		✓		✓		✓	✓
		20	4th formant delta standard deviation	✓	✓		✓	✓	✓	✓
		21	5th formant delta standard deviation	✓			✓	✓		✓
		22	6th formant delta standard deviation	✓			✓	✓		✓
	MFCC	23	7th MFCC minimum	✓			✓	✓		✓
		24	8th MFCC range	✓		✓	✓	✓		✓
		25	1st MFCC delta range			✓		✓	✓	✓
		26	2nd MFCC delta average	✓		✓	✓	✓	✓	✓
		27	7th MFCC delta minimum	✓		✓		✓		✓
		28	7th MFCC delta range	✓		✓		✓		✓
		29	8th MFCC delta average	✓		✓		✓		✓
		30	8th MFCC delta range	✓		✓		✓		✓
		31	9th MFCC delta average			✓		✓	✓	✓
Eye activity	Left-eye	1	average left-right movement	✓	✓		✓			✓
		2	maximum blink duration	✓		✓		✓		✓
		3	maximum duration of looking down		✓		✓		✓	✓
		4	maximum duration of looking left	✓		✓		✓		✓
	Right-eye	5	average left-right movement	✓		✓		✓		✓
		6	maximum blink duration	✓		✓		✓		✓
		7	maximum duration of looking down		✓	✓		✓		✓
Head pose		1	average roll angle			✓		✓	✓	✓
		2	average duration head direction left	✓	✓	✓		✓		
		3	avg. dur. fast movement for head direction left	✓	✓	✓	✓			
		4	average speed of up head direction		✓	✓	✓	✓		
		5	avg. dur. of continues rolling head anticlockwise		✓	✓	✓	✓		

Table 7.3: List of fixed features that exceed the T-statistic for the majority of dataset combinations

Variable set of features exceed the t-statistic: (based on the combined training data) Similar to the experiments on the in individual datasets, features that

exceed the t-statistic are identified from the combined training data and selected on the testing data. Therefore, the selected features might vary with each combination of datasets. However, in this approach, with leave-one-out cross-validation, common features that exceed the t-statistic in all turns are selected. That is, using the training subjects in each turn, we apply a T-test to all extracted functional features, then only those that commonly exceed the t-statistic in every turn are selected, similar to mutual ETF on individual datasets. Then, these common features are fixed and used for all leave-one-out cross-validation turns in the testing. On the other hand, in the train-test classification method, the features that exceed the t-statistic in the training set are selected on the testing set.

Fixed set of features exceed the t-statistic Unlike the variable set of features mentioned above, we seek to find a fixed set of features that commonly exceed the t-statistic on all individual datasets and combinations of the datasets. This fixed set of features is used on all individual and combinations of the datasets, not only to ensure a fair comparison between datasets combinations, but also to conclude a set of features that could generalise for the task of detecting depression. This set of features is selected based on the majority agreement of features that exceed the t-statistic on all individual datasets and combinations of the datasets (see Table 7.3 for the list of selected features).

Moreover, inspired by Schuller et al. [2010], corpus normalisation is used in this study to eliminate the differences of features from different dataset. Here, each dataset is normalised before its usage in combination with other corpora using Min-Max normalisation.

Results from combining datasets in the leave-one-subject-out method

Using different combinations of the three datasets, the proposed system is applied to validate the generalisability of the system in leave-one-out cross-validation, and the results are presented in Table 7.4.

Comparing classification results from individual datasets (see Table 7.2) with the current classification results (see Table 7.4), none of the dataset combination results improved compared with the results of their individual datasets. A reduction or at least no improvement from classification results when using different combinations of datasets compared to individual datasets was expected, given the several differences between the datasets. However, the classification results were statistically above chance level for most modalities with only two exceptions, which I believe support the generalisability claim of the proposed system and the selected features.

Several combinations of the three datasets have been used for classifying severely depressed subjects from low depressed or control subjects in order to identify which combination of datasets generalises better than the others. In general, the classification results of dataset combinations in the leave-one-out method are high, performing on average at 74% AR, which implies that this method had the ability to generalise

Modality/ Dataset	BlackDog + AVEC	BlackDog + Pitt	AVEC + Pitt	All three Datasets
Feature Selection	Variable set of features			
Speech prosody	84.8 (37)	77.6 (80)	80.0 (26)	74.6 (43)
Eye activity	63.0 (11)	78.6 (18)	61.4 (7)	68.5 (12)
Head pose	62.0 (11)	63.3 (20)	65.7 (1)	51.5 (1)
Feature fusion	84.8 (59)	60.2 (118)	87.1 (34)	80.0 (56)
Hybrid fusion	85.9	82.7	78.6	74.6
Feature Selection	Fixed set of features			
Speech prosody	76.1 (31)	78.6 (31)	54.3 (31)	74.6 (31)
Eye activity	69.6 (7)	73.5 (7)	78.6 (7)	70.8 (7)
Head pose	67.4 (5)	68.4 (5)	61.4 (5)	66.2 (5)
Feature fusion	82.6 (43)	82.7 (43)	88.6 (43)	79.2 (43)
Hybrid fusion	78.3	80.6	68.6	77.7

Table 7.4: Correct classification results (AR) in % and number of selected features (in parentheses) of dataset combinations using leave-one-out cross-validation. The second section of the table used a fixed set of features listed in Table 7.3.

to different combinations of datasets. In most cases, the classification results of each modality for each dataset combination are comparable.

Comparing feature selection methods, the fixed set of features performed statistically better than when using the variable feature set in the eye activity and head pose modalities as well as feature fusion. That implies, for these modalities, when the features are selected based on the specific dataset combination (the variable set of features), they have a lower generalisation ability in the classification problem when combining two or more datasets than when using a fixed set of features. Moreover, since the fixed set of features is selected based on the majority of features that exceed the T-statistic in all individual and combined datasets, it has more generalisation power for the classification problem when combining two or more datasets than using the variable set of features.

However, the speech prosody modality and hybrid fusion final results were remarkably better when using the variable set of features in most cases. The reduction in speech prosody classification results when using the fixed feature set might be due to the fixed features lacking essential features that differentiate the two classification classes in the dataset combinations, which implies that choosing a better set of features could be beneficial for the speech prosody modality.

Although, the eye activity, head pose, and feature fusion benefit from the fixed feature set, the hybrid fusion classification results were better when using the variable feature set. That might indicate that individual modalities had more agreements on the subjects' mental state when using the variable set of features than when using fixed feature set. Nevertheless, when combining the three datasets, the classification results are better using the fixed feature set than using the variable one, which supports the hypothesis that the fixed feature set has more generalisation power for the classification problem than using the variable set.

Individual modalities (speech prosody, eye activity, and head pose) are investigated for their generalisability to detect severe depression in a cross-cultural context. The speech prosody modality is investigated using the two methods of feature se-

lection, where using the variable feature set performed statistically better than using the fixed one. Nevertheless, speech modality consistently had high performance with only one exception where its performance was at chance level. This exception occurred when combining the AVEC dataset with the Pitt dataset using the fixed feature set. I do not believe that this low classification result is due to language differences, as the AVEC dataset combined with BlackDog dataset resulted in a high performance despite their different languages. The same applies to signal quality differences. The only reasonable explanation is that the selected features might not have the ability to generalise for this combination, especially that when using the variable set of features, the classification result for this combination was high (80% AR).

On the other hand, eye activity and head pose modalities using the fixed feature set performed statistically better than using the variable feature set. Moreover, the classification results of dataset combinations in the leave-one-out cross-validation using a fixed feature set for the eye activity modality are consistently high. This finding supports, once again, the claim that eye activity has distinguishing characteristics to detect severe depression from low or no depression behaviour.

In most cases of dataset combinations, the head pose modality performed lower than the other modalities using both feature selection methods, yet the classification results were above chance level with one exception. This exception is obtained when combining all three datasets and using the variable set of features. The reason behind the chance level performance for this combination might be due to the only significant feature that has been selected in this combination not being significant for the BlackDog and AVEC datasets, and therefore it could not generalise when using the combination of the three datasets (see feature #1 in head pose section in Table 7.3). Nevertheless, the reasonable performance (above chance level) for the head pose modality implies that the head pose holds useful information for the separation of severely depressed behaviour from low or non-depressed behaviour.

To complete the last step of the proposed system, individual modalities (speech prosody, eye activity and head pose) are fused at feature fusion and hybrid fusion levels. Feature fusion improves the classification results compared to the individual modality results that it fuses in both feature selection methods with all datasets combinations except one. The exceptional combination is BlackDog + Pitt when using the variable feature set, where the classification result was slightly lower than the lowest modality (catastrophic fusion). That might be due to the combined features being less correlated than when used individually. Despite the exceptional case, the improvements in feature fusion results suggest that speech prosody, eye activity and head pose features are correlated and complement each other on the task of detecting depression.

Hybrid fusion employs early and late fusion by combining the decisions from individual modalities with decisions from feature fusion, which might increase the confidence level of the final decision (see Figure 6.5). In dataset combinations using the leave-one-out cross-validation, hybrid fusion statistically improves the lowest individual modality classification results that it fuses in all cases for both feature

selection methods. Moreover, the hybrid fusion results either slightly decrease or slightly increase from the highest individual modalities that it fuses in most cases. The exception for this is when using the AVEC + Pitt combination in both feature selection methods, where the classification results were catastrophic (significantly lower than the lowest result). That might be due to the varied agreement/disagreement for the same subject, and since there is an even number of votes (4 votes), the majority voting of the hybrid fusion treats equal votes as a logic AND, which could have an effect on the results. Regardless, the classification results of the hybrid fusion were not catastrophic, which gives a stronger confidence in the hybrid fusion final results compared to individual modalities.

In general, regardless of the feature selection method, the classification results on dataset combinations in leave-one-out performed considerably better, even with the dataset differences. I believe that is due to the classifier learning from varied observations from each dataset, which therefore reduces the effect of overfitting the model to specific observation conditions.

Results from combining datasets in the train-test method

Beside using the leave-one-out cross-validation method, the train-test method of dataset combinations for generalisation investigation is also used, where one or two datasets are used for training and the remaining dataset(s) for testing. The classification results of generalisation using the train-test method using both the variable and fixed set of features methods are illustrated in Table 7.5.

Training Dataset(s)	BlackDog	AVEC	Pitt	BlackDog + AVEC	BlackDog + Pitt	AVEC + Pitt
Testing Dataset(s)	AVEC + Pitt	BlackDog + Pitt	BlackDog + AVEC	Pitt	AVEC	BlackDog
Feature Selection	Variable set of features					
Speech prosody	50.0 (63)	39.8 (94)	52.2 (100)	55.3 (50)	50.0 (104)	58.3 (55)
Eye activity	50.0 (20)	45.9 (40)	44.6 (21)	26.3 (16)	40.6 (20)	55.0 (15)
Head pose	50.0 (11)	37.8 (33)	41.3 (37)	21.1 (19)	34.4 (3)	60.0 (2)
Feature fusion	60.0 (94)	56.1 (167)	46.7 (158)	52.6 (85)	31.3 (64)	60.0 (72)
Hybrid fusion	41.4	45.9	51.1	31.6	34.4	61.7
Feature Selection	Fixed set of features					
Speech prosody	51.4 (31)	60.2 (31)	59.8 (31)	68.4 (31)	46.9 (31)	58.3 (31)
Eye activity	55.7 (7)	45.9 (7)	52.2 (7)	55.3 (7)	34.4 (7)	60.0 (7)
Head pose	42.9 (5)	42.9 (5)	42.4 (5)	28.9 (5)	34.4 (5)	61.7 (5)
Feature fusion	68.6 (43)	52.0 (43)	56.5 (43)	57.9 (43)	53.1 (43)	70.0 (43)
Hybrid fusion	65.7	39.8	46.7	60.5	34.4	66.7

Table 7.5: Classification results (AR) in % and number of selected features (in parentheses) of dataset combinations using train-test method. The second section of the table used a fixed set of features listed in Table 7.3.

In general, the classification results when using one or two datasets for training and using the remaining dataset(s) for testing are mostly at or lower than chance level with a few exceptions. That is expected as, unlike in the leave-one-out cross-validation method with dataset combinations, the classifier on the train-test method is trained on observations of dataset(s) that contain certain characteristics of the

training dataset(s), which risks over-fitting. The over-fitting issue reduces the classifierâĂŹs ability to generalise to separate and different dataset observations (unseen data).

Comparing different train-test dataset combinations, the only combination that has a reasonably above chance level classification result is the AVEC + Pitt dataset combination used for training and the BlackDog dataset for testing in both feature selection methods. This finding could indicate that when using AVEC and Pitt datasets, the classifier is trained on varied observations where the model is able to generalise to the BlackDog dataset observations. These variations might be due to: (1) the classification problem for both AVEC and Pitt is to classify severe depression from low depression, and, therefore, the model is trained on wide depression ranges, which might reduce the effect of overfitting, (2) the number of females in the AVEC + Pitt combination is more than half the total number of subjects (47 females out of 70 subjects). It has been reported that women amplify their mood when depressed [Nolen-Hoeksema, 1987], and therefore, the AVEC + Pitt combination model is trained on easily distinguishable observations, or (3) simply the differences in recording conditions and collection procedures varied, which made it flexible to generalise to the BlackDog recording conditions.

Using the variable set of features performed considerably lower than using the fixed feature set in most cases of dataset combinations for both feature selection methods. This finding implies that selecting features that exceed the T-statistic on the training dataset(s) to the testing dataset(s) has less generalisation power than the fixed feature set, which is similar to the finding when using leave-one-out cross-validation. Moreover, I believe that a better selection of the feature set could increase the generalisability and therefore the classification results. However, to obtain such a refined feature set, bigger and more varied depression datasets are required.

Speech prosody, eye activity and head pose modalities are investigated individually for their ability to generalise to different train-test combinations of datasets. When using a fixed feature set, speech prosody performed above chance level in most dataset combinations with two exceptions. One of these exceptions is the use of the BlackDog dataset for training and the combination of the AVEC and Pitt datasets for testing. The opposite training-testing combination for this exception worked better, which supports the hypothesis that the combination of the AVEC and Pitt datasets has varied observations, which made it difficult for a model trained on BlackDog dataset to generalise to such variation. The second exception is the use of the combination of the BlackDog and Pitt datasets for training and AVEC for testing, where their opposite training-testing combination worked much better. This finding supports the previous finding about varied observations and might imply that the AVEC dataset has more varied observations compared to the BlackDog and Pitt datasets. That source of the variation of the AVEC dataset is not clear at this point. However, it could be the variety in recording environments, where the recordings were performed in several recording sites (see Section 3.2.3). The different recording environment might had an effect in increasing the variation in audio and video quality and background.

With the fixed feature set, eye activity classification results were higher than the classification results of the head pose modality in all train-test datasets combinations except for one combination. Even though the classification results of both nonverbal modalities were at chance level in most train-test combination cases, there were three exceptions for the eye activity modality and one exception for the head pose modality where the results were above chance level. This finding could imply a high ability for eye activity to generalise to different datasets, which supports the claim that eye activity has distinguishing characteristics to detected depression. Moreover, this finding also implies that the head pose modality could hold useful information about depression behaviour.

As with previous classification problems, feature fusion and hybrid fusion were also investigated on the train-test dataset combinations. Feature fusion only improved the classification results from the highest classification result of individual modalities in a few cases of classifying different train-test dataset combinations for both feature selection methods. However, the classification results of feature fusion were not catastrophic. This finding suggests that individual modality features are correlated to the classification problem but not to each other, and therefore complement each other.

The only exception for this is using the BlackDog and Pitt datasets for training and AVEC for testing case when using the variable set of features, where the same training-testing combination had a remarkable improvement in feature fusion when using the fixed feature set. This exception could support the hypothesis that a refined feature set could be beneficial for generalising to several different datasets. On the other hand, hybrid fusion led to no improvement on classification results from the individual modalities that it fuses in most train-test combination classifications (only one exception), yet its results were not catastrophic (with the exception of two cases). The catastrophic cases in hybrid fusion indicate a disagreement about the subjects' mental state between the fused modalities. Moreover, having no improvement for the hybrid fusion is expected as most of the results are at chance level. Therefore, the modalities agreement on a given subject could vary. Nevertheless, hybrid fusion by using majority voting of the fused modalities could increase the confidence level of the final decision.

To sum up, generalising using the train-test method for classification of dataset combinations performed very low compared to the leave-one-out method, which might be caused by the overfitting problem, as the model is trained on specific conditions that prevent it from generalising to different observations.

By investigating the generalisability of the proposed system to different dataset combinations using the leave-one-out cross-validation and train-test methods for classification, a conclusion could be derived that when the classifier is trained on varied observations, the effect of overfitting, which is the main obstacle for cross-dataset generalisation, could be reduced, and therefore, the model could have better flexibility to generalise to new observations than when the classifier is trained on specific observations. That is, the more variability in the training observations, the better the generalisability to the testing observations. Moreover, a refined feature set

would be beneficial for the generalisation problem. However, such a refined feature set needs more observations and more advanced techniques than the ones used in this preliminary study.

7.4 Summary

In order to validate the usability and generalisability of the proposed system based on the previous chapters' investigations, this chapter investigated applying the proposed system on two other depression datasets. Beside the BlackDog dataset, AVEC, which is a self-rated depression dataset of German subjects, and Pitt, which is a clinically validated interview depression dataset of American subjects, were used in this chapter for the generalisation investigation.

These three datasets differ in several aspects including: collection procedure and task, depression diagnosis test and scale, cultural and language background, and recording environment. To reduce the dataset differences: (1) similar tasks of the collection procedure in each dataset were selected, which contain spontaneous self-directed speech, (2) the classification problem was categorised as a binary problem (i.e. severe depressed vs. low depressed/healthy controls), (3) functional features were extracted over the entire duration of each subject's segment to reduce duration variability, and (4) normalisation of the extracted features was applied to reduce recording environment and setting differences.

The proposed system: (1) investigates verbal (speech style and speech prosody) and nonverbal (eye activity and head pose) modalities, (2) extracts functional features from the verbal and nonverbal modalities over the entire subjects' segments, (3) pre- and post-normalises the extracted features, (4) selects features using T-test, (5) classifies depression in a binary manner (i.e. severe depressed vs. healthy controls), and finally (6) fuses the individual modalities.

However, several adjustments have been made to the proposed system to cope with the datasets differences, especially on modality preparations, such as: (1) excluding speech style modality as there is no equivalent for the AVEC dataset, (2) using different methods to separate speakers for speech prosody feature extraction, (3) using different numbers of annotated images to build subject-specific eye-AAM model, for the eye activity feature extraction, and (4) using different methods of face detection for the head pose feature extraction.

In order to test the generalisability of the proposed system, the system was applied on the three datasets individually and with different dataset combinations. Moreover, the combined datasets classifications were done in two ways: leave-one-subject-out cross-validation and train-test method, where one or two datasets are used for training and the remaining dataset(s) for testing.

Applying the proposed system on individual datasets as well as dataset combinations in leave-one-subject-out cross-validation gave high classification results even with the differences between the datasets. On the other hand, when using train-test dataset combinations, the results were at chance level in most cases, unless the

training dataset combinations had varied observations. With these findings, a conclusion can be made about the generalisability of the proposed system. That is, the proposed system has the ability to generalise to different datasets when applied individually, and is more likely to generalise to different dataset combinations given that the combination has varied observations, which would give the training model the flexibility to generalise to the testing observations. A model trained on varied observations reduces the effect of overfitting issues, which is the main obstacle for generalisation. Moreover, the more variability on the training observations, the better the generalisability to the testing observations.

Two methods of feature selection were investigated with the dataset combinations: (1) using a variable set of features selected based on the training dataset(s), and (2) using a fixed feature set based on the majority of features that exceed the T-statistic on all individual and combined datasets. The classification results were better when using the fixed feature set in most dataset combinations for both leave-one-subject-out cross-validation and train-test methods. This finding implies that the fixed feature set has generalisation power for the classification problem. Moreover, I believe that a better selection of the feature set could increase the generalisability and therefore the classification results. However, to obtain such a refined feature set, bigger and more varied depression datasets are required.

Individual modalities have been investigated for their generalisability to different datasets. Speech prosody features consistently resulted in high performance on both individual and combined datasets (for both leave-one-subject-out cross-validation and train-test methods), especially when using the fixed feature set for the combined datasets. This finding supports the generalisability of the speech prosody modality on different datasets, and implies that the extracted and normalised features are robust to different recording conditions. Similar to the speech prosody modality, the eye activity modality let to high classification results for both individual datasets as well as combined datasets with leave-one-subject-out cross-validation and some cases of train-test combinations. This finding supports the hypothesis that the eye activity modality has a strong ability to generalise to different datasets, which supports the claim that eye activity has distinguishing characteristics to detect depression. On the other hand, the head pose modality had reasonable classification results for both individual datasets as well as combined datasets with the leave-one-subject-out cross-validation. That could imply that the head pose holds useful information for the separation of severely depressed behaviour from low or non-depressed behaviour.

Finally, feature and hybrid fusion methods were completed as a final step of the proposed system for the generalisation investigation. In most cases of individual datasets as well as the combined datasets (for both leave-one-subject-out cross-validation and train-test methods), feature fusion improved the results of the individual modalities that it fuses. That implies that speech prosody, eye activity, and head pose modalities are correlated to the classification problem but not to each other and therefore complement each other on the task of detecting depression. On the other hand, hybrid fusion using majority voting on decisions from individual modalities and decisions from feature fusions, did not improve the classification results, yet the

results were not catastrophic. Hybrid fusion results are an indication of modality agreements, which increases the confidence level of the final decision.

Chapter 8

Conclusions

Depression is a common mental disorder that affects an estimated 350 million people worldwide and is therefore considered a burden not only on a personal and social level, but also on a economic level. This project is embedded in a larger, long-term research study that has the ultimate goal of developing an objective multimodal system that supports clinicians during the diagnosis and monitoring of clinical depression. The long-term research project is an interdisciplinary project across the ANU, University of Canberra, University of New South Wales and Queensland Institute of Medical Research that brings together computer scientists, speech scientists, psychologists, psychiatrists and neuroscientists.

In this dissertation, behavioural patterns that differentiate depressed patients from healthy controls were analysed in a clinical interview context. The thesisâĂŹ major research questions are to analyse verbal and nonverbal cues that characterise depression, as well as to investigate fusion techniques and the generalisability of the findings on different datasets. As stated in Chapter 1, these are examined with consideration to the results from the experimental work investigated in this thesis and summarised in Section 8.1. Section 8.2 states the contributions of this dissertation before discussing open issues and future directions in Section 8.3.

8.1 Research Questions

Q1. What are the distinguishing characteristics and most accurate configurations of verbal based depression detection in terms of extracted features, classification methods, and gender-dependence?

Investigating speech characteristics of depressed subjects compared to control subjects resulted in interesting findings as discussed in Chapter 4. Several speech style features were found to be statistically significant, possibly caused by higher cognitive load, as well as less involvement and less positive reaction in depressed patients. Speech prosody features were analysed statistically as well, and indicated a reduced control of vocal cords in speech production in depressed subjects.

Experiments were conducted to investigate the most accurate configurations of verbal based depression detection. While investigating extracted speech features, statistical functional features performed better (in terms of accuracy) than low-level

features. Moreover, prosody features were also analysed individually. Shimmer and formants consistently gave high classification results regardless of the feature type (i.e. low-level or functionals) and regardless of the classifier used (i.e. GMM, hybrid GMM-SVM, and SVM). Comparing feature selection methods, mutual features that exceed the T-statistic in all leave-one-out cross-validations resulted in the highest average recall (AR) classification results, which supports the hypothesis that T-test threshold can be used for feature selection in a binary classification task (depressed vs. non-depressed).

Q2. What are the specific nonverbal behaviour, movement, or activity patterns of the eyes and the head that could distinguish depressed patients from healthy control subjects in the classification task of detecting depression?

Eye activity and head pose movement patterns were investigated for their discriminative power for recognising depression from video data of subject interviews as described in Chapter 5. Analysing functional features of the eye activity modality found that the average distance between the eyelids was significantly smaller in between blinks and the average duration of blinks was significantly longer in depressed subjects, which might be an indication of fatigue and eye contact avoidance. In general, it could be concluded that eye movement abnormality is a physical cue as well as behavioural one, which is in line with the psychology literature in that depression leads to psychomotor retardation.

Statistical analyses on head pose and movement behavioural patterns found several distinguishing features: (1) depressed subjects had slower head movements, which may indicate fatigue, (2) the duration of looking to the right was longer in depressed patients, which might be an indicator of avoiding eye contact with the interviewer, (3) that overall change of head position in depression sufferers was significantly less than in healthy controls, and (4) the average duration of looking down was longer in depressed individuals, which could be an indicator of avoiding eye contact with the interviewer.

Q3. Which fusion method of the examined modalities (speech, eye, and head behavioural patterns) would improve the robustness and increase the accuracy of the depression recognition?

Fusing verbal and nonverbal temporal patterns of depression was investigated in Chapter 6, hypothesising an improvement in the fused result. Several methods of feature, score, decision, and hybrid fusions were investigated, as well as classifier fusion. The investigation found that hybrid fusion gave the highest classification results as well as a higher confidence level in the final results. That might be because hybrid fusion employs both early and late fusion techniques. Fusing classifiers' decisions was investigated, but did not lead for improved results, because the classifiers rarely agree with each other. Such inconsistency needs further investigation and might indicate a need for a better structure and modelling as well as parameter choices for the classifiers.

Q4. Are the findings and methods data-specific, or would they generalise to give similar results when used on different datasets of different recording environments as well as different

cultures and languages?

In order to validate the usability and generalisability of a proposed system based on previous investigations on the BlackDog dataset, applying the proposed system on two other depression datasets, namely the Pitt and AVEC datasets, was investigated in Chapter 7.

The proposed system: (1) investigates verbal (speech style and speech prosody) and nonverbal (eye activity and head pose) modalities, (2) extracts functional features from verbal and nonverbal modalities over the entire subjects' segments, (3) pre- and post-normalises the extracted features, (4) selects features using the T-statistic, (5) classifies depression in a binary manner (i.e. severely depressed vs. healthy controls), and finally (6) fuses the individual modalities.

The BlackDog, Pitt and AVEC datasets differ in several aspects including: collection procedure, depression diagnosis test and scale, cultural and language background, and recording environment. Several adjustments were made to the proposed system to cope with the dataset differences, especially on modality preparations.

Applying the proposed system on individual datasets as well as dataset combinations gave high classification results even with the differences between the datasets. This implies that the proposed system has the ability to generalise to different datasets, given that the combination has varied observations, which would give the training model the flexibility to generalise to the testing observations.

8.2 Summary of Contributions

Developing a system to detect depression objectively is a relatively new area of research, with many opportunities for contributions. Not only is there very little prior work on the automatic detection of depression from either single modality (audio or video) or fused modalities in the literature as reviewed in Section 2.3, these also used different methods, different measures, and were applied to different datasets, which makes the comparison of results with the results of my study even harder. Therefore, the comprehensive investigations of each step of developing a system to detect depression performed in my research allowed for a fair comparison of the results. This assisted not only with proposing a multimodal system to detect depression, but also with identifying the strengths and weaknesses that could be taken into account for future work as explained in Section 8.3. Moreover, the key contributions of this study are:

Expression-dependent: Positive and negative expression classification were investigated by using parts of the interview questions that were assumed to elicit the expressions in question. Even though the negative expression segment duration is longer than the duration of the positive expression segments, positive expression performed higher than negative expression in both verbal and nonverbal modalities. This finding implies similarity between depressed and controls while expressing negative emotions, and a higher difference between the two groups while expressing positive emotions. The high recognition of

depression using positive emotions could conclude that positive emotions are expressed less often in depressed subjects, and that negative emotions have less discriminatory power than positive emotions in detecting depression.

Thin slicing theory: For both gender-dependent and expression-dependent classifications, the sample size and number of observations were reduced substantially. Yet, recognition rates in both verbal and nonverbal modalities remain consistent. That could be explained by the psychological thin-slicing theory (as explained in Section 4.4). This theory supports the finding and the flexibility and robustness of the system to sample size reduction.

Eye activity: Eye activity features were extracted and used for depression classification in several contexts: general classification, gender-dependent classification, expression-dependent classification, generalisation across datasets classification, as well as individual features classification and when fused with other modalities. In all cases, the eye activity modality consistently gave good classification results to differentiate severely depressed subjects from low or non-depressed subjects. Moreover, fusing the eye activity features with other features of other modalities always contributes in improving classification results. This finding implies that eye activity is a strong characteristic to differentiate severe depressed from low or non-depressed behaviour.

Cross-dataset generalisation: Individual and fused modalities were investigated for their generalisability on different datasets. Speech prosody features consistently resulted in high performance on both individual and combined datasets. This finding supports the generalisability of the speech prosody modality on different datasets, as well as implies that the extracted and normalised features are robust to recording conditions. Similar to the speech prosody modality, the eye activity modality led to high classification results for both individual datasets as well as combined datasets. This finding supports the hypothesis that the eye activity modality has a strong ability to generalise on different datasets, which also supports the claim that eye activity has distinguishing characteristics to detect depression. On the other hand, the head pose modality gave a reasonable classification results for both individual datasets as well as combined datasets. That could imply that the head pose modality holds useful information about severely depressed behaviour, compared to low or non-depressed behaviour. Moreover, feature fusion improved the results of the individual modalities that it fuses. That indicates that speech prosody, eye activity, and head pose modalities are correlated and complement each other on the task of detecting depression. Even though hybrid fusion in the generalisation investigation did not improve the classification results, it increased the confidence level of the final decision. That is because hybrid fusion is an indication of modality agreements.

8.3 Future Work

Future research opportunities and areas of enhancement based on the investigations of this study are presented in this section.

Data Collection

Based on this study, several aspects could be considered for improving the collection of a depression dataset for the task of automatic depression detection. Although it is a common problem in similar studies, a known limitation in this work is the relatively small number of (depressed and control) subjects. As data collection at the Black Dog Institute is ongoing, reporting on a larger dataset in the future is anticipated. Having a larger number of subjects would allow to validate the investigated approaches and allow for using more advanced techniques as discussed in the following sections. In this work, only a subset of each dataset was used for the analysis because of the selection criteria applicable to the research questions, and to ensure a clinically valid balanced subset was used. However, future work could investigate using all subjects of these datasets to increase the sample size and validate the findings in this thesis.

In addition, future data collection could include personality assessments to investigate the effect of personality on detecting depression. Including personality assessments would add an extra dimension in understanding depression behaviour, which might increase the confidence of a system that detects depression. For example, a personality assessment would shed some light on understanding the subject's speech style or amount of gestures.

Furthermore, including different cultural backgrounds in the data collection would benefit cross-cultural investigations and increase the awareness of the effect of cultures on depression behaviour. It would be important to maintain the same paradigm in order to facilitate the comparison of findings. This thesis investigated depressed patients from broadly similar cultures. However, it has been reported that diverse cultures have different manifestations of depression and different cultural acceptance of depression. These differences may or may not affect an objective depression detection system, which will be interesting to investigate.

While this thesis investigated generalisation on two languages, English and German, and two dialects, Australian and American English, it would be informative to conduct generalisation tests across different languages. The investigation on different languages could shed some light on the physical characteristics of the speech production in depressed patients. While the English and German languages are stress-timed languages, an investigation of syllable-timed (e.g., Spanish, French, Italian) or mora-timed (e.g., Japanese) languages could be a future direction for depression detection.

This work investigated detecting depression based on expression-dependent content, where a clear distinction was found in the positive expression in depressed patients when compared to healthy controls. Therefore, including a longer duration of positive expressions in the data collection paradigm could be beneficial to obtain a more accurate depression detection, especially from the speech modalities.

Additionally, overall body movement was not investigated in this work as the recorded videos do not include a full body view. However, body movement has been shown to distinguish depressed patients from healthy controls. Including camera(s) to capture a full body view in the data acquisition would therefore enrich a future depression behaviour investigation.

Investigated Modalities

This work investigated verbal and nonverbal cues from speech style, speech prosody, eye activity, and head pose only. However, other cues could be investigated in future work. The linguistic modality, including word choices and sentence structure, could be a rich source of cues to detect depression. Moreover, adding facial expression and body movement modalities could enhance the depression detection accuracy and confidence level.

Furthermore, to get as accurate extracted features as possible, this work used manual annotation for speech, eye and head (automatic feature extraction was not the focus of this study). Speech annotation and speaker separation could be performed automatically using advanced speaker diarisation techniques. Regarding eye activity features, automated algorithms that measure blink and iris movements could be utilised for this task. For head pose and movement, a general face tracker could be effective in extracting head pose features. The same applies for linguistic, facial expression and body movement modalities. Therefore, having a fully automated system to extract and analyse features might be feasible for the task of detecting depression.

Advanced Fusion and Feature Selection Techniques

Since a larger dataset was not available for this work, weighted and complex fusion approaches could not be investigated. When a larger depression dataset is available, future work should include such fusion approaches, as the hypothesis is that they would increase the accuracy and the confidence level in the final decision.

The same applies for feature selection methods. Based on the generalisation investigation in Chapter 7, the selected features used gave reasonably good classification results. However, I believe that a better selection of a feature set could increase the objectivity and generalisability of a system that detects depression. Therefore, once a larger depression dataset is available, such investigations could be performed to obtain a refined feature set.

Classification Problem

Throughout this work, detecting depression was investigated in a binary classification manner (i.e. severe depressed vs. low depressed/healthy controls). However, with the inclusion of low depressed subjects, the classification problem could be further divided into three classes: severe depressed, low depressed, and healthy controls. Such a division could reduce confusion in differentiating low depressed

patients from healthy controls. Moreover, such a division could increase the generalisability and flexibility of a depression detection system to different depression datasets with varied depression severity.

Furthermore, having a regression classification problem to detect depression severity could be a next step for an advanced depression diagnosis system. Such a regression problem needs a large dataset with a variety of depression severity scores, noting the difficulty of obtaining an agreed severity score from clinical assessment. While the AVEC dataset was used here for investigations into the binary classification (severe vs. low depressed), I note the recent AVEC 2014 challenge focused on depression severity estimation and further work in this direction would be a natural extension of the research presented here.

Finally, investigations of whether gender-dependent classification influences the recognition rate were conducted in this study, where subjects of both gender were divided manually. Gender-dependent classification had a noticeable influence on recognising depression from verbal and nonverbal modalities, where females had a higher depression recognition rate than males, which is in line with previous gender difference studies. Given the higher classification results were obtained when using a gender-dependent model to classify depression than when using a gender-independent model in this study, it might be beneficial to model genders separately, if the size of the dataset allows it.

The current work has investigated a few classifiers, both discriminative (SVM and MLP) and generative (GMM and HFS), to achieve a binary classification. Besides considering a regression approach to estimate depression severity (as mentioned above), using different classifiers that provide conceptual structure (i.e. gender-dependent and expression-dependent) and account for different confidence levels is an interesting future direction. For example, adopting a Bayesian oriented framework to represent the output of each modality as likelihoods, where BayesâĂŹ rule would provide guidance for the fusion method and the weighting for these modalities, is one approach for future direction.

Appendix A

The following table shows the correct classification (✓) and the misclassification (✗) for each subject for each modality using an SVM classifier, in reference to Section 6.3.

Subject #	State	Gender	Speech Style	Speech Prosody	Eye Activity	Head Pose	Feature-Fusion	Hybrid: one-level
1	Depressed	Male	✓	✓	✓	✓	✓	✓
2			✓	✓	✓	✓	✓	✓
3			✓	✓	✓	✗	✓	✓
4			✗	✗	✓	✓	✗	✗
5			✗	✗	✓	✗	✗	✗
6			✓	✓	✗	✓	✓	✓
7			✓	✗	✓	✓	✗	✓
8			✗	✓	✓	✓	✗	✓
9			✓	✗	✓	✗	✓	✓
10			✓	✓	✗	✓	✗	✓
11			✓	✗	✓	✓	✓	✓
12			✓	✓	✓	✓	✓	✓
13			✗	✗	✓	✓	✓	✓
14			✓	✓	✗	✓	✓	✓
15			✓	✓	✓	✗	✓	✓
16		Female	✗	✓	✓	✗	✓	✓
17			✓	✓	✓	✗	✓	✓
18			✓	✓	✗	✓	✓	✓
19			✓	✓	✓	✓	✓	✓
20			✓	✓	✓	✓	✓	✓
21			✓	✓	✓	✓	✓	✓
22			✓	✓	✗	✓	✓	✓
23			✓	✓	✓	✓	✓	✓
24			✓	✓	✓	✓	✓	✓
25			✓	✓	✓	✓	✓	✓
26			✓	✓	✓	✓	✓	✓
27			✓	✓	✓	✓	✓	✓
28			✓	✓	✓	✓	✓	✓
29			✓	✓	✓	✓	✓	✓
30			✓	✓	✓	✓	✓	✓
31	Control	Male	✓	✓	✓	✓	✓	✓
32			✓	✗	✗	✗	✗	✗
33			✓	✓	✓	✓	✓	✓
34			✓	✓	✓	✗	✓	✓
35			✓	✗	✓	✓	✓	✓
36			✓	✓	✓	✓	✓	✓
37			✓	✓	✗	✗	✓	✓
38			✗	✓	✓	✓	✓	✓
39			✓	✓	✓	✗	✓	✓
40			✓	✓	✓	✓	✓	✓
41			✓	✓	✓	✓	✓	✓
42			✓	✓	✓	✓	✓	✓
43			✗	✓	✗	✓	✗	✗
44			✓	✓	✓	✗	✓	✓
45			✓	✓	✓	✓	✓	✓
46		Female	✓	✗	✓	✗	✗	✗
47			✓	✓	✗	✓	✓	✓
48			✓	✓	✓	✓	✓	✓
49			✗	✓	✓	✓	✓	✓
50			✓	✓	✓	✗	✓	✓
51			✓	✓	✓	✓	✓	✓
52			✓	✓	✓	✓	✓	✓
53			✓	✓	✓	✗	✓	✓
54			✓	✓	✓	✗	✓	✓
55			✓	✓	✓	✓	✓	✓
56			✗	✓	✓	✓	✓	✓
57			✓	✓	✓	✓	✓	✓
58			✓	✓	✓	✓	✓	✓
59			✓	✓	✓	✓	✓	✓
60			✓	✗	✓	✓	✓	✓

Bibliography

Abboud, B.; Davoine, F.; and Dang, M., 2004. Facial expression recognition and synthesis based on an appearance model. *Signal Processing: Image Communication*, 19, 8 (2004), 723 – 740. doi:http://dx.doi.org/10.1016/j.image.2004.05.009. http://www.sciencedirect.com/science/article/pii/S0923596504000475. (cited on page 22)

Abel, L.; Friedman, L.; Jesberger, J.; Malki, A.; and Meltzer, H., 1991. Quantitative assessment of smooth pursuit gain and catch-up saccades in schizophrenia and affective disorders. *Biological Psychiatry*, 29, 11 (1991), 1063 – 1072. doi:http://dx.doi.org/10.1016/0006-3223(91)90248-K. http://www.sciencedirect.com/science/article/pii/000632239190248K. (cited on pages 12 and 97)

Albrecht, A. T. and Herrick, C., 2010. *100 Questions & Answers About Depression*. Jones & Bartlett Learning, Sudbury, MA, 2 edn. ISBN 978-0-7637-4567-7. http://books.google.com.au/books?id=R4macqG9SzMC. (cited on pages 2, 7, and 8)

Alghowinem, S.; AlShehri, M.; Goecke, R.; and Wagner, M., 2014. Exploring eye activity as an indication of emotional states using an eye-tracking sensor. In *Intelligent Systems for Science and Information* (Eds. L. Chen; S. Kapoor; and R. Bhatia), vol. 542 of *Studies in Computational Intelligence*, 261–276. Springer International Publishing. ISBN 978-3-319-04701-0. doi:10.1007/978-3-319-04702-7_15. http://dx.doi.org/10.1007/978-3-319-04702-7_15. (cited on page 24)

Alghowinem, S.; Goecke, R.; Wagner, M.; Epps, J.; Breakspear, M.; and Parker, G., 2012. From joyous to clinically depressed: Mood detection using spontaneous speech. In *FLAIRS Conference*, 141–146. AAAI Press. http://dblp.uni-trier.de/db/conf/flairs/flairs2012.html#AlghowinemGWEBP12. (cited on page 78)

Alghowinem, S.; Goecke, R.; Wagner, M.; Epps, J.; Parker, G.; and Breakspear, M., 2013. Characterising depressed speech for classification. In *INTERSPEECH*, 2534–2538. ISCA. http://dblp.uni-trier.de/db/conf/interspeech/interspeech2013.html#AlghowinemGWEPB13. (cited on page 78)

Ambady, N. and Rosenthal, R., 1992. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, 111, 2 (1992), 256–274. doi:10.1037/0033-2909.111.2.256. (cited on pages 88 and 108)

American Psychiatric Association, 1994. *Diagnostic and Statistical Manual of Mental Disorders*, vol. 4th. American Psychiatric Association. (cited on pages 2 and 8)

Atrey, P.; Hossain, M.; El Saddik, A.; and Kankanhalli, M., 2010. Multimodal fusion for multimedia analysis: A survey. *Multimedia Systems*, 16, 6 (2010), 345–379. doi:10.1007/s00530-010-0182-0. http://dx.doi.org/10.1007/s00530-010-0182-0. (cited on pages 29 and 30)

Australian Bureau of Statistics, ABS, 2008. *Causes of death 2006*. 3303.0. ABS: Canberra. (cited on page 2)

Bacivarov, I.; Ionita, M.; and Corcoran, P., 2008. Statistical models of appearance for eye tracking and eye-blink detection and measurement. *IEEE Transactions on Consumer Electronics*, 54, 3 (August 2008), 1312–1320. doi:10.1109/TCE.2008.4637622. (cited on page 91)

Baik, S.-Y.; Bowers, B. J.; Oakley, L. D.; and Susman, J. L., 2005. The Recognition of Depression: The Primary Care Clinician's Perspective. *The Annals of Family Medicine*, 3, 1 (2005), 31–37. (cited on page 2)

Batliner, A.; Fischer, K.; Huber, R.; Spilker, J.; and Nöth, E., 2003. How to find trouble in communication. *Speech Communication*, 40, 1-2 (Apr. 2003), 117–143. doi:10.1016/S0167-6393(02)00079-1. http://dx.doi.org/10.1016/S0167-6393(02)00079-1. (cited on page 23)

Batliner, A. and Huber, R., 2007. Speaker characteristics and emotion classification. In *Speaker Classification I* (Ed. C. Müller), vol. 4343 of *Lecture Notes in Computer Science*, 138–151. Springer Berlin Heidelberg. ISBN 978-3-540-74186-2. doi:10.1007/978-3-540-74200-5_7. http://dx.doi.org/10.1007/978-3-540-74200-5_7. (cited on page 23)

Beck, A. T.; Steer, R. A.; Ball, R.; and Ranieri, W., 1996. Comparison of Beck Depression Inventories -IA and -II in psychiatric outpatients. *Journal of personality assessment*, 67, 3 (Dec 1996), 588–597. (cited on pages 2, 8, and 62)

Ben Mahmoud, H.; Ketata, R.; Ben Romdhane, T.; and Ben Ahmed, S., 2011. Hierarchical fuzzy signatures approach for a piloted quality management system. In *8th International Multi-Conference on Systems, Signals and Devices (SSD)*, 1–6. doi:10.1109/SSD.2011.5767379. (cited on page 28)

Berrios, G. E., 1985. The psychopathology of affectivity: Conceptual and historical aspects. *Psychological Medicine*, 15, 4 (Nov 1985), 745–758. (cited on page 14)

Bitouk, D.; Verma, R.; and Nenkova, A., 2010. Class-level spectral features for emotion recognition. *Speech Communication*, 52, 7-8 (2010), 613 – 625. doi:http://dx.doi.org/10.1016/j.specom.2010.02.010. http://www.sciencedirect.com/science/article/pii/S0167639310000348. (cited on page 81)

Boersma, P. and Weenink, D., 2009. Praat: Doing phonetics by computer. http://www.praat.org/. (cited on page 69)

Bylsma, L. M.; Morris, B. H.; and Rottenberg, J., 2008. A meta-analysis of emotional reactivity in major depressive disorder. *Clinical Psychology Review*, 28, 4 (2008), 676 – 691. doi:http://dx.doi.org/10.1016/j.cpr.2007.10.001. http://www.sciencedirect.com/science/article/pii/S0272735807001626. (cited on pages 8, 77, 86, 100, and 108)

Calvo, R. A. and D'Mello, S. K. D., 2011. *New Perspectives on Affect and Learning Technologies*, 3–10. Springer New York. http://www.springerlink.com/index/10.1007/978-1-4419-9625-1. (cited on page 14)

Caridakis, G.; Castellano, G.; Kessous, L.; Raouzaiou, A.; Malatesta, L.; Asteriadis, S.; and Karpouzis, K., 2007. Multimodal emotion recognition from expressive faces, body gestures and speech. In *Artificial Intelligence and Innovations 2007: from Theory to Applications* (Eds. C. Boukis; A. Pnevmatikakis; and L. Polymenakos), vol. 247 of *IFIP The International Federation for Information Processing*, 375–388. Springer US. ISBN 978-0-387-74160-4. doi:10.1007/978-0-387-74161-1_41. http://dx.doi.org/10.1007/978-0-387-74161-1_41. (cited on pages 15 and 29)

Castellano, G.; Villalba, S.; and Camurri, A., 2007. Recognising human emotions from body movement and gesture dynamics. In *Affective Computing and Intelligent Interaction* (Eds. A. Paiva; R. Prada; and R. Picard), vol. 4738 of *Lecture Notes in Computer Science*, 71–82. Springer Berlin Heidelberg. ISBN 978-3-540-74888-5. doi:10.1007/978-3-540-74889-2_7. http://dx.doi.org/10.1007/978-3-540-74889-2_7. (cited on page 24)

Chang, C. C. and Lin, C. J., 2001. Libsvm: a library for svm. *2006-03-04]. http:// www. csic. ntu. edu. tw/rcjlin/papers/lib. svm*, (2001). (cited on page 55)

Clavel, C.; Vasilescu, I.; Devillers, L.; Richard, G.; and Ehrette, T., 2008. Fear-type emotion recognition for future audio-based surveillance systems. *Speech Communication*, 50, 6 (Jun. 2008), 487–503. doi:10.1016/j.specom.2008.03.012. http://dx.doi.org/10.1016/j.specom.2008.03.012. (cited on page 19)

Cohn, J.; Kruez, T.; Matthews, I.; Yang, Y.; Nguyen, M. H.; Padilla, M.; Zhou, F.; and De la Torre, F., 2009. Detecting depression from facial actions and vocal prosody. In *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops ACII 2009*, 1–7. doi:10.1109/ACII.2009.5349358. (cited on pages 3, 13, 37, 39, 40, and 43)

Crawford, T. J.; Haeger, B.; Kennard, C.; Reveley, M. A.; and Henderson, L., 1995. Saccadic abnormalities in psychotic patients. I. Neuroleptic-free psychotic patients. *Psychological Medicine*, 25, 3 (May 1995), 461–471. (cited on pages 13 and 97)

Cummins, N.; Epps, J.; Breakspear, M.; and Goecke, R., 2011. An investigation of depressed speech detection: Features and normalization. In *INTERSPEECH*, 2997–3000. ISCA. http://dblp.uni-trier.de/db/conf/interspeech/interspeech2011.html#CumminsEBG11. (cited on pages 37 and 43)

Cummins, N.; Joshi, J.; Dhall, A.; Sethu, V.; Goecke, R.; and Epps, J., 2013. Diagnosis of depression by behavioural signals: A multimodal approach. In *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, AVEC '13 (Barcelona, Spain, 2013), 11–20. ACM, New York, NY, USA. doi:10.1145/2512530.2512535. http://doi.acm.org/10.1145/2512530.2512535. (cited on pages 40 and 44)

Dasarathy, B., 1997. Sensor fusion potential exploitation-innovative architectures and illustrative applications. *Proceedings of the IEEE*, 85, 1 (Jan 1997), 24–38. doi:10.1109/5.554206. (cited on page 30)

Datcu, D. and Rothkrantz, L. J. M., 2008. Semantic audio-visual data fusion for automatic emotion recognition. In *Euromedia'2008 Porto*, 58–65. Eurosis, Ghent. http://mmi.tudelft.nl/pub/dragos/_datcu_euromedia08.pdf. (cited on page 27)

de Jong, N. and Wempe, T., 2009. Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, 41, 2 (2009), 385–390. doi:10.3758/BRM.41.2.385. http://dx.doi.org/10.3758/BRM.41.2.385. (cited on page 70)

de Krom, G., 1993. A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals. *Journal of Speech, Language, and Hearing Research*, 36, 2 (1993), 254–266. (cited on page 11)

De Silva, P. R. and Bianchi-Berthouze, N., 2004. Modeling human affective postures: An information theoretic characterization of posture features. *Computer Animation and Virtual Worlds*, 15, 3-4 (2004), 269–276. doi:10.1002/cav.29. http://dx.doi.org/10.1002/cav.29. (cited on page 24)

Dementhon, D. and Davis, L., 1995. Model-based object pose in 25 lines of code. *International Journal of Computer Vision*, 15, 1-2 (1995), 123–141. doi:10.1007/BF01450852. http://dx.doi.org/10.1007/BF01450852. (cited on pages 101 and 102)

Depression Scores Conversion, Online. Inventory of Depressive Symptomatology (IDS) & Quick Inventory of Depressive Symptomatology (QIDS). http://www.ids-qids.org/index2.html. (cited on page 129)

Devins, G. M. and Orme, C. M., 1985. Center for epidemiologic studies depression scale. *Test critiques*, 2 (1985), 144–60. (cited on page 35)

Dittmann, A., 1987. Body movements as diagnostic cues in affective disorders. *Depression and expressive behavior*, (1987). (cited on page 13)

D'Mello, S. and Graesser, A., 2010. Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. *User Modeling and User-Adapted Interaction*, 20, 2 (2010), 147–187. doi:10.1007/s11257-010-9074-4. http://dx.doi.org/10.1007/s11257-010-9074-4. (cited on page 15)

D'Mello, S. and Kory, J., 2012. Consistent but modest: A meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, ICMI '12 (Santa Monica, California, USA, 2012), 31–38. ACM, New York, NY, USA. doi:10.1145/2388676.2388686. http://doi.acm.org/10.1145/2388676.2388686. (cited on pages 28 and 29)

Drucker, H.; Burges, C. J.; Kaufman, L.; Smola, A.; and Vapnik, V., 1997. Support vector regression machines. *Advances in neural information processing systems*, 9 (1997), 155–161. (cited on pages 27 and 41)

Eisenberg, N. and Spinrad, T. L., 2004. Emotion-related regulation: Sharpening the definition. *Child Development*, 75, 2 (2004), 334–339. doi:10.1111/j.1467-8624.2004.00674.x. http://dx.doi.org/10.1111/j.1467-8624.2004.00674.x. (cited on page 95)

Ekman, P., 1994. Moods Emotions And Traits. In *P. Ekman & R. Davidson (Eds.) The Nature of Emotion: Fundamental Questions*, 15–19. Oxford University Press, New York. (cited on pages 8, 13, 72, 77, 86, and 108)

Ekman, P., 1999. Basic emotions. In *John Wiley & Sons, The Handbook of Cognition and Emotion*, 45–60. (cited on pages 9, 14, and 15)

Ekman, P. and Davidson, R. J., 1994. Affective science: A research agenda. *The nature of emotion: Fundamental questions*, (1994), 411–430. (cited on page 9)

Ekman, P. and Fridlund, A. J., 1987. Assesment Of Facial Behvior In Affective Disorders. In *Depression and Expressive Behavior*, 37–56. Lawrence Erlbaum, Hillsdale, N.J. (cited on pages 8, 13, 72, 77, 86, and 108)

Ekman, P. and Friesen, W. V., 1972. Hand movements. *Journal of Communication*, 22, 4 (1972), 353–374. doi:10.1111/j.1460-2466.1972.tb00163.x. http://dx.doi.org/10.1111/j.1460-2466.1972.tb00163.x. (cited on page 13)

Ekman, P. and Friesen, W. V., 1974. Nonverbal behavior and psychopathology. *The psychology of depression Contemporary theory and research*, (1974), 203–232. (cited on page 13)

Ekman, P.; Matsumoto, D.; and Friesen, W. V., 1997a. Facial expression in affective disorders. In *What the Face Reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding Systems (FACS)* (Eds. P. Ekman and E. Rosenberg). Oxford. (cited on pages 9 and 16)

Ekman, P.; Matsumoto, D.; and Friesen, W. V., 1997b. Facial expression in affective disorders. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*, 2 (1997). (cited on pages 13 and 23)

El Ayadi, M.; Kamel, M. S.; and Karray, F., 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*,

44, 3 (Mar. 2011), 572–587. doi:10.1016/j.patcog.2010.09.020. http://dx.doi.org/10.1016/j.patcog.2010.09.020. (cited on page 27)

Ellgring, H., 1989. *Non-verbal communication in depression*. Cambridge University Press. (cited on pages 12 and 13)

Ellgring, H. and Scherer, K., 1996. Vocal indicators of mood change in depression. *Journal of Nonverbal Behavior*, 20, 2 (1996), 83–110. doi:10.1007/BF02253071. http://dx.doi.org/10.1007/BF02253071. (cited on pages 9, 11, 72, 73, and 79)

Eyben, F.; Wöllmer, M.; and Schuller, B., 2010. Opensmile: The Munich versatile and fast open-source audio feature extractor. In *Proceedings of the International Conference on Multimedia*, MM '10 (Firenze, Italy, 2010), 1459–1462. ACM, New York, NY, USA. doi:10.1145/1873951.1874246. http://doi.acm.org/10.1145/1873951.1874246. (cited on page 78)

Fasel, B. and Luettin, J., 2003. Automatic facial expression analysis: a survey. *Pattern Recognition*, 36, 1 (2003), 259 – 275. doi:http://dx.doi.org/10.1016/S0031-3203(02)00052-3. http://www.sciencedirect.com/science/article/pii/S0031320302000523. (cited on page 23)

Flint, A. J.; Black, S. E.; Campbell-Taylor, I.; Gailey, G. F.; and Levinton, C., 1993. Abnormal speech articulation, psychomotor retardation, and subcortical dysfunction in major depression. *Journal of Psychiatric Research*, 27, 3 (1993), 309 – 319. doi:http://dx.doi.org/10.1016/0022-3956(93)90041-Y. http://www.sciencedirect.com/science/article/pii/002239569390041Y. (cited on pages 10 and 81)

Fossi, L.; Faravelli, C.; and Paoli, M., 1984. The ethological approach to the assessment of depressive disorders. *The Journal of nervous and mental disease*, 172, 6 (Jun 1984), 332–341. (cited on pages 12, 13, and 105)

Fournier, J. C.; DeRubeis, R. J.; Hollon, S. D.; Dimidjian, S.; Amsterdam, J. D.; Shelton, R. C.; and Fawcett, J., 2010. Antidepressant drug effects and depression severity. *JAMA: The Journal of the American Medical Association*, 303, 1 (2010), 47–53. (cited on page 61)

France, D.; Shiavi, R.; Silverman, S.; Silverman, M.; and Wilkes, D., 2000. Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE Transactions on Biomedical Engineering*, 47, 7 (July 2000), 829–837. doi:10.1109/10.846676. (cited on pages 10, 37, 41, 43, and 51)

France, J., 2001. Depression and other mood disorders. *Communication and mental illness: Theoretical and practical approaches*, (2001), 65. (cited on page 13)

Frank, M. G. and Ekman, P., 1996. Physiological Effects Of The Smile. *Directions in Psychiatry*, 16, 25 (1996), 1–8. (cited on page 16)

Gee, A. and Cipolla, R., 1994. Determining the gaze of faces in images. *Image and Vision Computing*, 12, 10 (1994), 639 – 647. doi:http://dx.doi.org/10.1016/0262-8856(94)90039-6. http://www.sciencedirect.com/science/article/pii/0262885694900396. (cited on page 101)

Goldman-Eisler, F., 1968. *Psycholinguistics: Experiments in spontaneous speech*. Academic Press, London and New York. (cited on page 71)

Grandin, T., 1995. How people with autism think. In *Learning and Cognition in Autism* (Eds. E. Schopler and G. Mesibov), Current Issues in Autism, 137–156. Springer US. ISBN 978-1-4899-1288-6. doi:10.1007/978-1-4899-1286-2_8. http://dx.doi.org/10.1007/978-1-4899-1286-2_8. (cited on page 95)

Gross, J. J. and Levenson, R. W., 1995. Emotion elicitation using films. *Cognition & Emotion*, 9, 1 (1995), 87–108. doi:10.1080/02699939508408966. http://dx.doi.org/10.1080/02699939508408966. (cited on page 59)

Grubbs, F., 1969. Procedures for detecting outlying observations in samples. *Technometrics*, 11, 1 (1969), 1–21. doi:10.1080/00401706.1969.10490657. http://www.tandfonline.com/doi/abs/10.1080/00401706.1969.10490657. (cited on pages 94 and 103)

Gruber, J.; Oveis, C.; Keltner, D.; and Johnson, S. L., 2011. A discrete emotions approach to positive emotion disturbance in depression. *Cognition and Emotion*, 25, 1 (2011), 40–52. doi:10.1080/02699931003615984. http://dx.doi.org/10.1080/02699931003615984. (cited on page 72)

Gunes, H. and Pantic, M., 2010. Automatic, Dimensional and Continuous Emotion Recognition. *International Journal of Synthetic Emotions (IJSE)*, 1, 1 (jan 2010), 68–99. doi:doi:10.4018/jse.2010101605. (cited on pages 16, 27, and 29)

Gunes, H. and Piccardi, M., 2005. Fusing face and body gesture for machine recognition of emotions. In *IEEE International Workshop on Robot and Human Interactive Communication, ROMAN 2005.*, 306–311. doi:10.1109/ROMAN.2005.1513796. (cited on page 24)

Gunes, H. and Piccardi, M., 2009. Automatic temporal segment detection and affect recognition from face and body display. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 39, 1 (Feb 2009), 64–84. doi:10.1109/TSMCB.2008.927269. (cited on pages 21 and 24)

Gunes, H.; Schuller, B.; Pantic, M.; and Cowie, R., 2011. Emotion representation, analysis and synthesis in continuous space: A survey. In *IEEE International Conference on Automatic Face Gesture Recognition and Workshops (FG 2011)*, 827–834. doi:10.1109/FG.2011.5771357. (cited on page 14)

Guyon, I. and Elisseeff, A., 2003. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3 (Mar. 2003), 1157–1182. http://dl.acm.org/citation.cfm?id=944919.944968. (cited on pages 25 and 26)

Guyon, I.; Gunn, S.; Nikravesh, M.; and Zadeh, L. A., 2006. *Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing)*. Studies in Fuzziness and Soft Computing, Physica-Verlag. Springer-Verlag New York, Inc., Secaucus, NJ, USA. ISBN 3540354875. (cited on page 26)

Guze, S. B. and Robins, E., 1970. Suicide and Primary Affective Disorders. *The British Journal of Psychiatry*, 117, 539 (oct 1970), 437–438. (cited on page 1)

Hager, G. D. and Toyama, K., 1998. X Vision: A Portable Substrate for Real-Time Vision Applications. *Computer Vision and Image Understanding*, 69, 1 (Jan 1998), 23–37. doi:10.1006/cviu.1997.0586. http://dx.doi.org/10.1006/cviu.1997.0586. (cited on page 90)

Hale III, W. W.; Jansen, J. H.; Bouhuys, A. L.; Jenner, J. A.; and van den Hoofdakker, R. H., 1997. Non-verbal behavioral interactions of depressed patients with partners and strangers: The role of behavioral social support and involvement in depression persistence. *Journal of affective disorders*, 44, 2-3 (Jul 1997), 111–122. (cited on pages 12, 72, and 105)

Hall, M. A. and Smith, L. A., 1998. Practical feature subset selection for machine learning. *Computer Science*, 98 (1998), 181âĂŞ191. http://researchcommons.waikato.ac.nz/handle/10289/1512. (cited on page 25)

Hamilton, M., 1960. A rating scale for depression. *Journal of neurology, neurosurgery, and psychiatry*, 23, 1 (1960), 56–62. (cited on pages 2 and 8)

Hansen, D. W. and Pece, A. E., 2005. Eye tracking in the wild. *Computer Vision and Image Understanding*, 98, 1 (2005), 155 – 181. doi:http://dx.doi.org/10.1016/j.cviu.2004.07.013. http://www.sciencedirect.com/science/article/pii/S107731420400116X. Special Issue on Eye Detection and Tracking. (cited on pages 22 and 90)

Heinzel, G.; Rüdiger, A.; Schilling, R.; and Hannover, T., 2002. Spectrum and spectral density estimation by the discrete fourier transform (dft), including a comprehensive list of window functions and some new flat-top windows. *Max Planck Institute*, 12 (2002), 122. (cited on page 19)

Heylen, D., 2006. Head gestures, gaze and the principles of conversational structure. *International Journal of Humanoid Robotics*, 3, 03 (2006), 241–267. doi:10.1142/S0219843606000746. http://www.worldscientific.com/doi/abs/10.1142/S0219843606000746. (cited on page 12)

Hoch, S.; Althoff, F.; McGlaun, G.; and Rigoll, G., 2005. Bimodal fusion of emotional data in an automotive environment. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005 (ICASSP '05)*, vol. 2, ii/1085–ii/1088 Vol. 2. doi:10.1109/ICASSP.2005.1415597. (cited on page 17)

Hollon, S. D.; DeRubeis, R. J.; Evans, M. D.; Wiemer, M. J.; Garvey, M. J.; Grove, W. M.; and Tuason, V. B., 1992. Cognitive therapy and pharmacotherapy for

depression: Singly and in combination. *Archives of General Psychiatry*, 49, 10 (1992), 774–781. doi:10.1001/archpsyc.1992.01820100018004. http://dx.doi.org/10.1001/archpsyc.1992.01820100018004. (cited on page 33)

Horprasert, T.; Yacoob, Y.; and Davis, L. S., 1997. Computing 3d head orientation from a monocular image sequence. In *25th Annual AIPR Workshop on Emerging Applications of Computer Vision*, vol. 2962, 244–252. International Society for Optics and Photonics (SPIE). doi:10.1117/12.267830. http://dx.doi.org/10.1117/12.267830. (cited on page 101)

Hussain, M.; Monkaresi, H.; and Calvo, R., 2012. Categorical vs. dimensional representations in multimodal affect detection during learning. In *Intelligent Tutoring Systems* (Eds. S. Cerri; W. Clancey; G. Papadourakis; and K. Panourgia), vol. 7315 of *Lecture Notes in Computer Science*, 78–83. Springer Berlin Heidelberg. ISBN 978-3-642-30949-6. doi:10.1007/978-3-642-30950-2_11. http://dx.doi.org/10.1007/978-3-642-30950-2_11. (cited on page 15)

Ioannou, S. V.; Raouzaiou, A. T.; Tzouvaras, V. A.; Mailis, T. P.; Karpouzis, K. C.; and Kollias, S. D., 2005. Emotion recognition through facial expression analysis based on a neurofuzzy network. *Neural Networks*, 18, 4 (2005), 423 – 435. doi:http://dx.doi.org/10.1016/j.neunet.2005.03.004. http://www.sciencedirect.com/science/article/pii/S0893608005000377. Emotion and Brain. (cited on page 24)

Jackson, M., 1983. Knowledge of the body. *Man*, 18, 2 (1983), 327–345. http://www.jstor.org/stable/2801438. (cited on page 13)

Jaimes, A. and Sebe, N., 2007. Multimodal human-computer interaction: A survey. *Computer Vision and Image Understanding*, 108, 1-2 (2007), 116 – 134. doi:http://dx.doi.org/10.1016/j.cviu.2006.10.019. http://www.sciencedirect.com/science/article/pii/S1077314206002335. Special Issue on Vision for Human-Computer Interaction. (cited on pages 14 and 15)

Jain, A.; Nandakumar, K.; and Ross, A., 2005. Score normalization in multimodal biometric systems. *Pattern Recognition*, 38, 12 (2005), 2270 – 2285. doi:http://dx.doi.org/10.1016/j.patcog.2005.01.012. http://www.sciencedirect.com/science/article/pii/S0031320305000592. (cited on page 41)

Jayalakshmi, T. and Santhakumaran, A., 2011. Statistical normalization and back propagation for classification. *International Journal of Computer Theory and Engineering*, 3, 1 (2011), 89–93. (cited on page 17)

Jiang, B.; Valstar, M.; and Pantic, M., 2011. Action unit detection using sparse appearance descriptors in space-time video volumes. In *IEEE International Conference on Automatic Face Gesture Recognition and Workshops (FG 2011)*, 314–321. doi:10.1109/FG.2011.5771416. (cited on page 23)

Jones, I. H. and Pansa, M., 1979. Some nonverbal aspects of depression and schizophrenia occurring during the interview. *The Journal of nervous and mental disease*, 167, 7 (Jul 1979), 402–409. (cited on page 13)

Joshi, J.; Dhall, A.; Goecke, R.; Breakspear, M.; and Parker, G., 2012. Neural-net classification for spatio-temporal descriptor based depression analysis. In *21st International Conference on Pattern Recognition (ICPR)*, 2634–2638. (cited on pages 39 and 43)

Joshi, J.; Dhall, A.; Goecke, R.; and Cohn, J., 2013a. Relative body parts movement for automatic depression analysis. In *Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*, 492–497. doi:10.1109/ACII.2013.87. (cited on pages 40 and 43)

Joshi, J.; Goecke, R.; Alghowinem, S.; Dhall, A.; Wagner, M.; Epps, J.; Parker, G.; and Breakspear, M., 2013b. Multimodal assistive technologies for depression diagnosis and monitoring. *Journal on Multimodal User Interfaces*, 7, 3 (2013), 217–228. doi:10.1007/s12193-013-0123-2. http://dx.doi.org/10.1007/s12193-013-0123-2. (cited on pages 41 and 43)

Joshi, J.; Goecke, R.; Parker, G.; and Breakspear, M., 2013c. Can body expressions contribute to automatic depression analysis? In *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 1–7. doi:10.1109/FG.2013.6553796. (cited on pages 39 and 43)

Kammoun, M. and Ellouze, N., 2006. Spectral features detection of speech emotion and speaking styles recognition based on hmm classifier. In *Proceedings of the 5th WSEAS International Conference on Signal Processing*, SIP'06 (Istanbul, Turkey, 2006), 1–5. World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, Wisconsin, USA. http://dl.acm.org/citation.cfm?id=1983937.1983939. (cited on page 23)

Kanade, T.; Cohn, J.; and Tian, Y., 2000. Comprehensive database for facial expression analysis. In *Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, 46–53. doi:10.1109/AFGR.2000.840611. (cited on page 16)

Kathmann, N.; Hochrein, A.; Uwer, R.; and Bondy, B., 2003. Deficits in gain of smooth pursuit eye movements in schizophrenia and affective disorder patients and their unaffected relatives. *The American Journal of Psychiatry*, 160, 4 (Apr 2003), 696–702. http://www.ncbi.nlm.nih.gov/pubmed/12668358. (cited on page 12)

Kiloh, L. G.; Andrews, G.; and Neilson, M., 1988. The long-term outcome of depressive illness. *The British Journal of Psychiatry*, 153, DEC (dec 1988), 752–757. (cited on page 2)

Kleinsmith, A. and Bianchi-Berthouze, N., 2013. Affective body expression perception and recognition: A survey. *IEEE Transactions on Affective Computing*, 4, 1 (Jan 2013), 15–33. doi:10.1109/T-AFFC.2012.16. (cited on page 24)

Kleinsmith, A.; Bianchi-Berthouze, N.; and Steed, A., 2011. Automatic recognition of non-acted affective postures. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 41, 4 (Aug 2011), 1027–1038. doi:10.1109/TSMCB.2010.2103557. (cited on page 24)

Koolagudi, S. and Rao, K., 2012. Emotion recognition from speech: A review. *International Journal of Speech Technology*, 15, 2 (2012), 99–117. doi:10.1007/s10772-011-9125-1. http://dx.doi.org/10.1007/s10772-011-9125-1. (cited on page 10)

Kroenke, K. and Spitzer, R. L., 2002. The PHQ-9: A new depression diagnostic and severity measure. *Psychiatric Annals*, 32, 9 (2002), 1–7. (cited on page 8)

Kuny, S. and Stassen, H. H., 1993. Speaking behavior and voice sound characteristics in depressive patients during recovery. *Journal of Psychiatric Research*, 27, 3 (1993), 289–307. http://www.ncbi.nlm.nih.gov/pubmed/8295161. (cited on pages 9 and 79)

Kupfer, D. and Foster, F., 1972. Interval between onset of sleep and rapid-eye-movement sleep as an indicator of depression. *The Lancet*, 300, 7779 (1972), 684 – 686. doi:http://dx.doi.org/10.1016/S0140-6736(72)92090-9. http://www.sciencedirect.com/science/article/pii/S0140673672920909. Originally published as Volume 2, Issue 7779. (cited on page 97)

Kwon, O.-w.; Chan, K.; Hao, J.; and Lee, T.-w., 2003. Emotion recognition by speech signals. *Eighth European Conference on Speech Communication and Technology*, (2003), 125–128. (cited on page 81)

Lacy, T. J. and McManis, S. E., 1994. Psychogenic stridor. *General Hospital Psychiatry*, 16, 3 (1994), 213 – 223. doi:http://dx.doi.org/10.1016/0163-8343(94)90104-X. http://www.sciencedirect.com/science/article/pii/016383439490104X. (cited on page 80)

Lanata, A.; Armato, A.; Valenza, G.; and Scilingo, E., 2011. Eye tracking and pupil size variation as response to affective stimuli: A preliminary study. In *5th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth)*, 78–84. (cited on page 24)

Lang, P. J.; Bradley, M. M.; and Cuthbert, B. N., 2005. *International affective picture system (IAPS): Affective ratings of pictures and instruction manual*. NIMH, Center for the Study of Emotion & Attention. (cited on page 59)

Laptev, I., 2005. On space-time interest points. *International Journal of Computer Vision*, 64, 2-3 (2005), 107–123. doi:10.1007/s11263-005-1838-7. http://dx.doi.org/10.1007/s11263-005-1838-7. (cited on page 22)

Lecrubier, Y., 2000. Depressive illness and disability. *European neuropsychopharmacology the journal of the European College of Neuropsychopharmacology*, 10 Suppl 4 (2000), S439–S443. http://www.ncbi.nlm.nih.gov/pubmed/11114489. (cited on page 1)

Lefter, I.; Rothkrantz, L.; Wiggers, P.; and van Leeuwen, D., 2010. Emotion recognition from speech by combining databases and fusion of classifiers. In *Text, Speech and Dialogue* (Eds. P. Sojka; A. Horak; I. KopeÃğek; and K. Pala), vol. 6231 of *Lecture Notes in Computer Science*, 353–360. Springer Berlin Heidelberg. ISBN 978-3-642-15759-2. doi:10.1007/978-3-642-15760-8_45. http://dx.doi.org/10.1007/978-3-642-15760-8_45. (cited on page 32)

Li, D.; Winfield, D.; and Parkhurst, D., 2005. Starburst: A hybrid algorithm for video-based eye tracking combining feature-based and model-based approaches. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops (CVPR Workshops)*, 79–79. doi:10.1109/CVPR.2005.531. (cited on pages 21, 22, 24, and 90)

Lipton, R. B.; Levin, S.; and Holzman, P. S., 1980. Horizontal and vertical pursuit eye movements, the oculocephalic reflex, and the functional psychoses. *Psychiatry Research*, 3, 2 (1980), 193–203. http://www.ncbi.nlm.nih.gov/pubmed/6947312. (cited on pages 12 and 97)

Low, L.-S.; Maddage, M.; Lech, M.; Sheeber, L.; and Allen, N., 2011. Detection of clinical depression in adolescents speech during family interactions. *IEEE Transactions on Biomedical Engineering*, 58, 3 (March 2011), 574–586. doi:10.1109/TBME.2010.2091640. (cited on pages 11, 38, 44, and 81)

Mackintosh, J. H.; Kumar, R.; and Kitamura, T., 1983. Blink rate in psychiatric illness. *The British Journal of Psychiatry*, 143, 1 (1983), 55–57. http://www.ncbi.nlm.nih.gov/pubmed/6882993. (cited on page 13)

Maddage, M.; Senaratne, R.; Low, L.-S.; Lech, M.; and Allen, N., 2009. Video-based detection of the clinical depression in adolescents. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 3723–3726. doi:10.1109/IEMBS.2009.5334815. (cited on pages 39 and 44)

Martin, A.; Doddington, G.; Kamm, T.; and Ordowski, M., 1997. The DET curve in assessment of detection task performance. *Proceedings Eurospeech*, (1997), 1895–1898. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.21.903&rep=rep1&type=ps. (cited on page 32)

Martins, P. and Batista, J., 2008. Monocular head pose estimation. In *Image Analysis and Recognition* (Eds. A. Campilho and M. Kamel), vol. 5112 of *Lecture Notes in Computer Science*, 357–368. Springer Berlin Heidelberg. ISBN 978-3-540-69811-1. doi:10.1007/978-3-540-69812-8_35. http://dx.doi.org/10.1007/978-3-540-69812-8_35. (cited on pages 101 and 103)

Mathers, C.; Boerma, J.; and Fat, D., 2008. *The Global Burden of Disease: 2004 Update*. WHO, Geneva, Switzerland. (cited on page 1)

McIntyre, G.; Goecke, R.; Hyett, M.; Green, M.; and Breakspear, M., 2009. An approach for automatically measuring facial activity in depressed subjects. In *3rd*

International Conference on Affective Computing and Intelligent Interaction and Workshops (ACII), 1–8. doi:10.1109/ACII.2009.5349593. (cited on page 13)

Mendis, B. and Gedeon, T., 2008. A comparison: Fuzzy signatures and choquet integral. In *IEEE International Conference on Fuzzy Systems, 2008. FUZZ-IEEE 2008. (IEEE World Congress on Computational Intelligence)*, 1464–1471. doi:10.1109/FUZZY.2008.4630565. (cited on page 28)

Mendis, B.; Gedeon, T.; and Koczy, L., 2006. Learning generalized weighted relevance aggregation operators using levenberg-marquardt method. In *Sixth International Conference on Hybrid Intelligent Systems (HIS '06)*, 34–34. doi:10.1109/HIS.2006.264917. (cited on page 121)

Meng, H.; Huang, D.; Wang, H.; Yang, H.; AI-Shuraifi, M.; and Wang, Y., 2013. Depression recognition based on dynamic facial and vocal expression features using partial least square regression. In *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, AVEC '13 (Barcelona, Spain, 2013), 21–30. ACM, New York, NY, USA. doi:10.1145/2512530.2512532. http://doi.acm.org/10.1145/2512530.2512532. (cited on pages 41 and 44)

Molina, L.; Belanche, L.; and Nebot, A., 2002. Feature selection algorithms: a survey and experimental evaluation. In *IEEE International Conference on Data Mining (ICDM)*, 306–313. doi:10.1109/ICDM.2002.1183917. (cited on page 25)

Monkaresi, H.; Hussain, M.; and Calvo, R., 2012. Classification of affects using head movement, skin color features and physiological signals. In *Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference on*, 2664–2669. doi:10.1109/ICSMC.2012.6378149. (cited on page 15)

Moore, E.; Clements, M.; Peifer, J.; and Weisser, L., 2003. Analysis of prosodic variation in speech for clinical depression. In *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 3, 2925–2928. doi:10.1109/IEMBS.2003.1280531. (cited on pages 37 and 43)

Moore, E.; Clements, M.; Peifer, J.; and Weisser, L., 2004. Comparing objective feature statistics of speech for classifying clinical depression. In *26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEMBS '04)*, vol. 1, 17–20. doi:10.1109/IEMBS.2004.1403079. (cited on pages 9, 11, 37, 43, 73, and 79)

Moore, E.; Clements, M.; Peifer, J.; and Weisser, L., 2008. Critical analysis of the impact of glottal features in the classification of clinical depression in speech. *IEEE Transactions on Biomedical Engineering*, 55, 1 (Jan 2008), 96–107. doi:10.1109/TBME.2007.900562. (cited on pages 10, 11, 37, 43, 73, 79, and 81)

Morris, T.; Blenkhorn, P.; and Zaidi, F., 2002. Blink detection for real-time eye tracking. *Journal of Network and Computer Applications*, 25, 2 (2002), 129 –

143. doi:http://dx.doi.org/10.1006/jnca.2002.0130. http://www.sciencedirect.com/science/article/pii/S108480450290130X. (cited on page 90)

Mundt, J. C.; Snyder, P. J.; Cannizzaro, M. S.; Chappie, K.; and Geralts, D. S., 2007. Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. *Journal of Neurolinguistics*, 20, 1 (2007), 50–64. doi:10.1016/j.jneuroling.2006.04.001. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3022333&tool=pmcentrez&rendertype=abstract. (cited on pages 2, 9, 33, 34, 36, 38, 43, and 79)

Murphy-Chutorian, E. and Trivedi, M., 2009. Head pose estimation in computer vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31, 4 (April 2009), 607–626. doi:10.1109/TPAMI.2008.106. (cited on page 101)

Murray, H. A., 1943. *Thematic apperception test.* Harvard University Press. ISBN 9780674877207. (cited on page 63)

NeuroCom, 2014. Head shake-sensory organization test (hs-sot). http://resourcesonbalance.com/neurocom/protocols/sensoryimpairment/hs-sot.aspx. (cited on page 101)

Niedermaier, N.; Bohrer, E.; Schulte, K.; Schlattmann, P.; and Heuser, I., 2004. Misdiagnosed patients with bipolar disorder: comorbidities, treatment patterns, and direct treatment costs. *The Journal of clinical psychiatry*, 66, 11 (2004), 1619–1623. (cited on page 2)

Nikolaidis, A. and Pitas, I., 2000. Facial feature extraction and pose determination. *Pattern Recognition*, 33, 11 (2000), 1783 – 1791. doi:http://dx.doi.org/10.1016/S0031-3203(99)00176-4. http://www.sciencedirect.com/science/article/pii/S0031320399001764. (cited on page 101)

Nilsonne, A., 1988. Speech characteristics as indicators of depressive illness. *Acta Psychiatrica Scandinavica*, 77, 3 (1988), 253–263. doi:10.1111/j.1600-0447.1988.tb05118.x. http://dx.doi.org/10.1111/j.1600-0447.1988.tb05118.x. (cited on pages 9 and 79)

Nolen-Hoeksema, S., 1987. Sex differences in unipolar depression: Evidence and theory. *Psychol Bull*, , 101 (Mar 1987), 259–282. (cited on pages 75, 84, 99, 107, 118, and 141)

Nunes, A.; Coimbra, L.; and Teixeira, A., 2010. Voice quality of European Portuguese emotional speech. In *Computational Processing of the Portuguese Language* (Eds. T. Pardo; A. Branco; A. Klautau; R. Vieira; and V. de Lima), vol. 6001 of *Lecture Notes in Computer Science*, 142–151. Springer Berlin Heidelberg. ISBN 978-3-642-12319-1. doi:10.1007/978-3-642-12320-7_19. http://dx.doi.org/10.1007/978-3-642-12320-7_19. (cited on pages 11 and 81)

Ojala, T.; Pietikäinen, M.; and Harwood, D., 1996. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29, 1 (1996), 51 – 59. doi:http://dx.doi.org/10.1016/0031-3203(95)00067-4. http://www.sciencedirect.com/science/article/pii/0031320395000674. (cited on page 22)

Ooi, K.; Lech, M.; and Allen, N., 2013. Multichannel weighted speech classification system for prediction of major depression in adolescents. *IEEE Transactions on Biomedical Engineering*, 60, 2 (Feb 2013), 497–506. doi:10.1109/TBME.2012.2228646. (cited on pages 38 and 44)

Ooi, K.; Low, L.; Lech, M.; and Allen, N., 2011. Prediction of clinical depression in adolescents using facial image analysis. In *12th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, 1–4. Delft, The Netherlands. (cited on pages 39 and 44)

Ooi, K.; Low, L.-S.; Lech, M.; and Allen, N., 2012. Early prediction of major depression in adolescents using glottal wave characteristics and teager energy parameters. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4613–4616. doi:10.1109/ICASSP.2012.6288946. (cited on pages 38 and 44)

Ozdas, A.; Shiavi, R.; Silverman, S.; Silverman, M.; and Wilkes, D., 2000. Analysis of fundamental frequency for near term suicidal risk assessment. In *IEEE International Conference on Systems, Man, and Cybernetics*, vol. 3, 1853–1858. doi:10.1109/ICSMC.2000.886379. (cited on page 10)

Ozdas, A.; Shiavi, R.; Silverman, S.; Silverman, M.; and Wilkes, D., 2004. Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk. *IEEE Transactions on Biomedical Engineering*, 51, 9 (Sept 2004), 1530–1540. doi:10.1109/TBME.2004.827544. (cited on pages 10, 37, 41, 43, 51, and 80)

Pantic, M. and Rothkrantz, L., 2000. Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22, 12 (Dec 2000), 1424–1445. doi:10.1109/34.895976. (cited on page 23)

Pantic, M. and Rothkrantz, L., 2003. Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*, 91, 9 (Sept 2003), 1370–1390. doi:10.1109/JPROC.2003.817122. (cited on pages 14 and 16)

Pantic, M.; Sebe, N.; Cohn, J. F.; and Huang, T., 2005. Affective multimodal human-computer interaction. In *Proceedings of the 13th Annual ACM International Conference on Multimedia*, MULTIMEDIA '05 (Hilton, Singapore, 2005), 669–676. ACM, New York, NY, USA. doi:10.1145/1101149.1101299. http://doi.acm.org/10.1145/1101149.1101299. (cited on page 14)

Parker, G. and Hadzi-Pavlovic, D., 1996. *Melancholia: A disorder of movement and mood: A phenomenological and neurobiological review*. Cambridge University Press. (cited on pages 72 and 105)

Parker, G.; Hadzi-Pavlovic, D.; Wilhelm, K.; Hickie, I.; Brodaty, H.; Boyce, P.; Mitchell, P.; and Eyers, K., 1994. Defining melancholia: Properties of a refined sign-based measure. *The British Journal of Psychiatry*, 164, 3 (Mar 1994), 316–326. (cited on pages 72 and 105)

Pedersen, J.; Schelde, J. T. M.; Hannibal, E.; Behnke, K.; Nielsen, B. M.; and Hertz, M., 1988. An ethological description of depression. *Acta Psychiatrica Scandinavica*, 78, 3 (1988), 320–330. doi:10.1111/j.1600-0447.1988.tb06343.x. http://dx.doi.org/10.1111/j.1600-0447.1988.tb06343.x. (cited on page 12)

Pence, B.; O'Donnell, J.; and Gaynes, B., 2012. The depression treatment cascade in primary care: A public health perspective. *Current Psychiatry Reports*, 14, 4 (2012), 328–335. doi:10.1007/s11920-012-0274-y. http://dx.doi.org/10.1007/s11920-012-0274-y. (cited on page 2)

Picard, R. W., 1997. *Affective Computing*. 321. MIT Press, Cambridge, MA, USA. ISBN 0262161702. doi:10.1007/BF01238028. http://vismod.media.mit.edu/tech-reports/TR-321.pdf. (cited on page 14)

Polzehl, T.; Sundaram, S.; Ketabdar, H.; Wagner, M.; and Metze, F., 2009. Emotion Classification in Children's Speech Using Fusion of Acoustic and Linguistic Features. In *Proceedings INTERSPEECH*, 340–343. Brighton. doi:10.1109/ICSC.2009.32. (cited on page 25)

Pope, B.; Blass, T.; Siegman, A. W.; and Raher, J., 1970. Anxiety and depression in speech. *Journal of Consulting and Clinical Psychology*, 35, 1 (1970), 128–133. http://www.ncbi.nlm.nih.gov/pubmed/5487600. (cited on pages 11 and 73)

Prendergast, M., 2006. *Understanding Depression*. Penguin Group Australia, VIC Australia. (cited on pages 2, 7, and 105)

Ranelli, C. J. and Miller, R. E., 1981. Behavioral predictors of amitriptyline response in depression. *The American journal of psychiatry*, 138, 1 (1981), 30–34. doi:10.1176/ajp.138.1.30. http://ajp.psychiatryonline.org/doi/abs/10.1176/ajp.138.1.30. (cited on page 13)

Rattani, A.; Kisku, D.; Bicego, M.; and Tistarelli, M., 2007. Feature level fusion of face and fingerprint biometrics. In *First IEEE International Conference on Biometrics: Theory, Applications, and Systems (BTAS)*, 1–6. doi:10.1109/BTAS.2007.4401919. (cited on page 29)

Reed, L. I., 2005. Timing characteristics of smiles in relation to history and symptomatology of depression. http://d-scholarship.pitt.edu/7345/. (cited on pages 11 and 72)

Rice, J. A. and Silverman, B. W., 1991. Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53, 1 (1991), pp. 233–243. http://www.jstor.org/stable/2345738. (cited on page 30)

Rodriguez, L.; Crespo, A.; Lara, M.; and Mezcua, B., 2008. Study of different fusion techniques for multimodal biometric authentication. In *IEEE International Conference on Wireless and Mobile Computing Networking and Communications (WIMOB '08)*, 666–671. doi:10.1109/WiMob.2008.29. (cited on page 30)

Ruchkin, V.; Sukhodolsky, D. G.; Vermeiren, R.; Koposov, R. A.; and Schwab-Stone, M., 2006. Depressive symptoms and associated psychopathology in urban adolescents: a cross-cultural study of three countries. *The Journal of nervous and mental disease*, 194, 2 (2006), 106–113. (cited on page 9)

Rush, A. J.; Trivedi, M. H.; Ibrahim, H. M.; Carmody, T. J.; Arnow, B.; Klein, D. N.; Markowitz, J. C.; Ninan, P. T.; Kornstein, S.; Manber, R.; et al., 2003. The 16-item quick inventory of depressive symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. *Biological psychiatry*, 54, 5 (2003), 573–583. (cited on pages 2 and 59)

Russell, J. A., 1979. Affective space is bipolar. *Journal of Personality and Social Psychology*, 37, 3 (1979), 345–356. (cited on page 14)

Sanchez, M. H.; Vergyri, D.; Ferrer, L.; Richey, C.; Garcia, P.; Knoth, B.; and Jarrold, W., 2011. Using prosodic and spectral features in detecting depression in elderly males. In *INTERSPEECH*, 3001–3004. ISCA. (cited on pages 38 and 44)

Saragih, J. and Goecke, R., 2006. Iterative error bound minimisation for aam alignment. In *18th International Conference on Pattern Recognition (ICPR)*, vol. 2, 1196–1195. doi:10.1109/ICPR.2006.730. (cited on pages 91 and 102)

Saragih, J.; Lucey, S.; and Cohn, J., 2009. Face alignment through subspace constrained mean-shifts. In *IEEE 12th International Conference on Computer Vision*, 1034–1041. doi:10.1109/ICCV.2009.5459377. (cited on pages 50 and 131)

Scherer, K. R., 1987. Vocal assessment of affective disorders. In *Depression and Expressive Behavior* (Ed. J. D. Maser), 57–82. Lawrence Erlbaum Associates, Hillsdale, USA. (cited on pages 10, 11, 80, and 81)

Scherer, S.; Stratou, G.; Gratch, J.; and Morency, L.-P., 2013a. Investigating voice quality as a speaker-independent indicator of depression and PTSD. In *Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Lyon, France. http://ict.usc.edu/pubs/Investigating_Voice_Quality_as_a_Speaker-Independent_Indicator_of_Depression_and_PTSD.pdf. (cited on pages 38 and 44)

Scherer, S.; Stratou, G.; Mahmoud, M.; Boberg, J.; Gratch, J.; Rizzo, A.; and Morency, L.-P., 2013b. Automatic behavior descriptors for psychological disorder analysis. In *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 1–8. doi:10.1109/FG.2013.6553789. (cited on pages 24 and 40)

Scherer, S.; Stratou, G.; and Morency, L.-P., 2013c. Audiovisual behavior descriptors for depression assessment. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ICMI '13 (Sydney, Australia, 2013), 135–140. ACM, New York, NY, USA. doi:10.1145/2522848.2522886. http://doi.acm.org/10.1145/2522848.2522886. (cited on pages 41 and 44)

Schlosberg, H., 1954. Three dimensions of emotion. *Psychological review*, 61, 2 (1954), 81–88. http://dx.doi.org/10.1037/h0054570. (cited on page 14)

Schuller, B.; Batliner, A.; Seppi, D.; Steidl, S.; Vogt, T.; Wagner, J.; Devillers, L.; Vidrascu, L.; Amir, N.; Kessous, L.; and Aharonson, V., 2007. The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals. In *Proceedings of INTERSPEECH*, 2253–2256. Antwerp, Belgium. (cited on page 23)

Schuller, B.; Batliner, A.; Steidl, S.; and Seppi, D., 2011a. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, 53, 9-10 (2011), 1062 – 1087. doi:http://dx.doi.org/10.1016/j.specom.2011.01.011. http://www.sciencedirect.com/science/article/pii/S0167639311000185. Sensing Emotion and Affect - Facing Realism in Speech Processing. (cited on pages 27, 28, 30, and 31)

Schuller, B.; Müller, R.; Lang, M. K.; Rigoll, G.; et al., 2005. Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles. In *INTERSPEECH*, 805–808. (cited on page 81)

Schuller, B.; Vlasenko, B.; Eyben, F.; Wollmer, M.; Stuhlsatz, A.; Wendemuth, A.; and Rigoll, G., 2010. Cross-corpus acoustic emotion recognition: Variances and strategies. *IEEE Transactions on Affective Computing*, 1, 2 (July 2010), 119–131. doi:10.1109/T-AFFC.2010.8. (cited on pages 32 and 137)

Schuller, B.; Zhang, Z.; Weninger, F.; and Rigoll, G., 2011b. Using multiple databases for training in emotion recognition: To unite or to vote? In *INTERSPEECH*, 1553–1556. (cited on page 32)

Sebe, N.; Cohen, I.; and Huang, T. S., 2005. Multimodal emotion recognition. *Handbook of Pattern Recognition and Computer Vision*, 4 (2005), 387–419. (cited on page 14)

Seo, K., 2004. Face pose estimation system by combining hybrid ica-svm learning and re-registration. In *Asian Conference on Computer Vision (ACCV)*, 27–30. (cited on page 101)

Shan, C.; Gong, S.; and Mcowan, P. W., 2007. Beyond Facial Expressions : Learning Human Emotion from Body Gestures. In *Proceedings of the British Machine Vision Conference*, 43.1–43.10. BMVA Press. doi:10.5244/C.21.43. (cited on page 24)

Sheehan, D. V.; Lecrubier, Y.; Sheehan, K. H.; Amorim, P.; Janavs, J.; Weiller, E.; Hergueta, T.; Baker, R.; and Dunbar, G. C., 1998. The mini-international neuropsychiatric interview (m.i.n.i.): The development and validation of a structured diagnostic psychiatric interview for dsm-iv and icd-10. In *Journal of Clinical Psychiatry*, vol. 59, 22–33. (cited on page 58)

Silverman, S. and Silverman, M., 2002. Methods and apparatus for evaluating near-term suicidal risk using vocal parameters. http://www.google.com/patents/US20020077825. US Patent App. 09/935,294. (cited on page 33)

Simon, G. E.; Katon, W.; Rutter, C.; VonKorff, M.; Lin, E.; Robinson, P.; Bush, T.; Walker, E. A.; Ludman, E.; and Russo, J., 1998. Impact of improved depression treatment in primary care on daily functioning and disability. *Psychological medicine*, 28, 3 (May 1998), 693–701. http://www.ncbi.nlm.nih.gov/pubmed/9626725. (cited on page 2)

Singer, K., 1975. Depressive disorders from a transcultural perspective. *Social Science & Medicine (1967)*, 9, 6 (1975), 289 – 301. doi:http://dx.doi.org/10.1016/0037-7856(75)90001-3. http://www.sciencedirect.com/science/article/pii/0037785675900013. (cited on page 9)

Sloman, A.; Chrisley, R.; and Scheutz, M., 2005. The Architectural Basis of Affective States and Processes. In *Compare A Journal Of Comparative Education* (Eds. J. M. Fellous and M. A. Arbib), vol. 4281, 203–244. Oxford University Press. http://eprints.sussex.ac.uk/1274/. (cited on page 14)

Sobin, C. and Sackeim, H. A., 1997. Psychomotor symptoms of depression. *American Journal of Psychiatry*, 154, 1 (1997), 4–17. (cited on pages 11, 13, and 72)

Song, Y.; Morency, L.-P.; and Davis, R., 2013. Learning a sparse codebook of facial and body microexpressions for emotion recognition. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ICMI '13 (Sydney, Australia, 2013), 237–244. ACM, New York, NY, USA. doi:10.1145/2522848.2522851. http://doi.acm.org/10.1145/2522848.2522851. (cited on page 23)

Stratou, G.; Scherer, S.; Gratch, J.; and Morency, L.-P., 2013. Automatic nonverbal behavior indicators of depression and ptsd: Exploring gender differences. In *Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*, 147–152. doi:10.1109/ACII.2013.31. (cited on pages 24, 39, 44, 99, and 107)

Sweeney, J. A.; Strojwas, M. H.; Mann, J. J.; and Thase, M. E., 1998. Prefrontal and cerebellar abnormalities in major depression: Evidence from oculomotor studies. 43, 8 (1998), 584–594. http://www.ncbi.nlm.nih.gov/pubmed/9564443. (cited on page 13)

Tamas, K. and Koczy, L., 2008. Selection from a fuzzy signature database by mamdani-algorithm. In *6th International Symposium on Applied Machine Intelligence and Informatics (SAMI)*, 63–68. doi:10.1109/SAMI.2008.4469135. (cited on page 28)

Tranter, S. and Reynolds, D., 2006. An overview of automatic speaker diarization systems. *IEEE Transactions on Audio, Speech, and Language Processing*, 14, 5 (Sept 2006), 1557–1565. doi:10.1109/TASL.2006.878256. (cited on pages 17, 49, and 68)

Trevino, A.; Quatieri, T.; and Malyska, N., 2011. Phonologically-based biomarkers for major depressive disorder. *EURASIP Journal on Advances in Signal Processing*, 2011, 1 (2011), 42. doi:10.1186/1687-6180-2011-42. http://dx.doi.org/10.1186/1687-6180-2011-42. (cited on pages 38 and 43)

Troisi, A. and Moles, A., 1999. Gender differences in depression:: an ethological study of nonverbal behavior during interviews. *Journal of Psychiatric Research*, 33, 3 (1999), 243 – 250. doi:http://dx.doi.org/10.1016/S0022-3956(98)00064-8. http://www.sciencedirect.com/science/article/pii/S0022395698000648. (cited on pages 75, 84, 99, and 107)

Truong, K. and Leeuwen, D. V., 2007. An 'open-set' detection evaluation methodology for automatic emotion recognition in speech. In *In ParaLing'07: Workshop on Paralinguistic Speech - between models and data*, 5–10. Saarbrucken, Germany. (cited on page 32)

Tsai, J. L. and Chentsova-Dutton, Y., 2002. Understanding depression across cultures. (2002). (cited on page 9)

Tseng, W.-S., 2001. Culture and mental health. In *Handbook of Cultural Psychiatry* (Ed. W.-S. Tseng), 123 –. Academic Press, San Diego. ISBN 978-0-12-701632-0. doi:http://dx.doi.org/10.1016/B978-012701632-0/50079-6. http://www.sciencedirect.com/science/article/pii/B9780127016320500796. (cited on page 9)

Tulyakov, S.; Jaeger, S.; Govindaraju, V.; and Doermann, D., 2008. Review of classifier combination methods. In *Machine Learning in Document Analysis and Recognition* (Eds. S. Marinai and H. Fujisawa), vol. 90 of *Studies in Computational Intelligence*, 361–386. Springer Berlin Heidelberg. ISBN 978-3-540-76279-9. doi:10.1007/978-3-540-76280-5_14. http://dx.doi.org/10.1007/978-3-540-76280-5_14. (cited on pages 29 and 30)

US Department of Health and Human Services, 2000. Healthy People 2010: Understanding and Improving Health. http://www.citeulike.org/user/mallinga/article/3892097. (cited on page 1)

Valstar, M.; Schuller, B.; Smith, K.; Eyben, F.; Jiang, B.; Bilakhia, S.; Schnieder, S.; Cowie, R.; and Pantic, M., 2013. Avec 2013: The continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, AVEC '13 (Barcelona, Spain, 2013), 3–10. ACM, New York, NY, USA. doi:10.1145/2512530.2512533. http://doi.acm.org/10.1145/2512530.2512533. (cited on pages 35 and 62)

Venegas, J. and Clark, E., 2011. Wechsler test of adult reading. In *Encyclopedia of Clinical Neuropsychology* (Eds. J. Kreutzer; J. DeLuca; and B. Caplan), 2693–2694.

Springer New York. ISBN 978-0-387-79947-6. doi:10.1007/978-0-387-79948-3_1500. http://dx.doi.org/10.1007/978-0-387-79948-3_1500. (cited on page 58)

Viola, P. and Jones, M., 2001. Rapid object detection using a boosted cascade of simple features. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, I–511–I–518. doi:10.1109/CVPR.2001.990517. (cited on pages 91 and 103)

Vogt, T. and Andre, E., 2005. Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In *IEEE International Conference on Multimedia and Expo (ICME)*, 474–477. doi:10.1109/ICME.2005.1521463. (cited on page 81)

Wagner, M. Multibiometric authentication. In *Advanced Topics in Biometrics*, chap. 17, 419–434. doi:10.1142/9789814287852_0017. http://www.worldscientific.com/doi/abs/10.1142/9789814287852_0017. (cited on page 17)

Wang, J.-G. and Sung, E., 2007. EM enhancement of 3D head pose estimated by point at infinity. *Image and Vision Computing*, 25, 12 (2007), 1864 – 1874. doi:http://dx.doi.org/10.1016/j.imavis.2005.12.017. http://www.sciencedirect.com/science/article/pii/S0262885606002848. The age of human computer interaction. (cited on page 101)

Waxer, P. H., 1974. Therapist training in nonverbal communication i: Nonverbal cues for depression. *Journal of Clinical Psychology*, 30, 2 (1974), 215–218. doi:10.1002/1097-4679(197404)30:2<215::AID-JCLP2270300229>3.0.CO;2-Q. http://dx.doi.org/10.1002/1097-4679(197404)30:2<215::AID-JCLP2270300229>3.0.CO;2-Q. (cited on pages 12 and 105)

Williamson, J. R.; Quatieri, T. F.; Helfer, B. S.; Horwitz, R.; Yu, B.; and Mehta, D. D., 2013. Vocal biomarkers of depression based on motor incoordination. In *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, AVEC '13 (Barcelona, Spain, 2013), 41–48. ACM, New York, NY, USA. doi:10.1145/2512530.2512531. http://doi.acm.org/10.1145/2512530.2512531. (cited on pages 39 and 44)

Wold, H., 2004. Partial least squares. In *Encyclopedia of Statistical Sciences*. John Wiley & Sons, Inc. ISBN 9780471667193. doi:10.1002/0471667196.ess1914.pub2. http://dx.doi.org/10.1002/0471667196.ess1914.pub2. (cited on page 41)

Wolfel, M. and McDonough, J., 2009. Speech feature extraction. In *Distant Speech Recognition*, 135–179. John Wiley & Sons, Ltd. ISBN 9780470714089. doi:10.1002/9780470714089.ch5. http://dx.doi.org/10.1002/9780470714089.ch5. (cited on page 19)

World Health Organization, W., 2003. *The world health report 2003: shaping the future*. World Health Organization. (cited on page 1)

World Health Organization, W., 2012. *World health statistics 2012*. World Health Organization. (cited on pages 1 and 2)

Wundt, W., 2009. Outlines of psychology (1897). In *Foundations of psychological thought: A history of psychology*, 36–44. Sage Publications, Inc. (cited on page 14)

Xu, G. and Huang, J. Z., 2012. Asymptotic optimality and efficient computation of the leave-subject-out cross-validation. *The Annals of Statistics*, 40, 6 (2012), 3003–3030. doi:10.1214/12-AOS1063. http://arxiv.org/abs/1302.4607. (cited on page 31)

Yacoob, Y. and Davis, L., 1996. Recognizing human facial expressions from long image sequences using optical flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18, 6 (Jun 1996), 636–642. doi:10.1109/34.506414. (cited on page 22)

Yang, G.; Lin, Y.; and Bhattacharya, P., 2005. A driver fatigue recognition model using fusion of multiple features. In *IEEE International Conference on Systems, Man and Cybernetics*, vol. 2, 1777–1784. doi:10.1109/ICSMC.2005.1571406. (cited on page 24)

Yang, M.-H.; Kriegman, D.; and Ahuja, N., 2002. Detecting faces in images: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24, 1 (Jan 2002), 34–58. doi:10.1109/34.982883. (cited on page 21)

Yang, Y.; Fairbairn, C.; and Cohn, J., 2013. Detecting depression severity from vocal prosody. *IEEE Transactions on Affective Computing*, 4, 2 (April 2013), 142–150. doi:10.1109/T-AFFC.2012.38. (cited on pages xvii, 49, and 61)

Young, L. and Sheena, D., 1975. Survey of eye movement recording methods. *Behavior Research Methods & Instrumentation*, 7, 5 (1975), 397–429. doi:10.3758/BF03201553. http://dx.doi.org/10.3758/BF03201553. (cited on page 90)

Zeng, Z.; Pantic, M.; Roisman, G.; and Huang, T., 2009. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31, 1 (Jan 2009), 39–58. doi:10.1109/TPAMI.2008.52. (cited on pages 15, 27, and 30)

Zhang, Z.; Lyons, M.; Schuster, M.; and Akamatsu, S., 1998. Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron. In *Third IEEE International Conference on Automatic Face and Gesture Recognition*, 454–459. doi:10.1109/AFGR.1998.670990. (cited on page 23)

Zhao, W.; Chellappa, R.; Phillips, P. J.; and Rosenfeld, A., 2003. Face recognition: A literature survey. *ACM Computing Surveys*, 35, 4 (Dec. 2003), 399–458. doi:10.1145/954339.954342. http://doi.acm.org/10.1145/954339.954342. (cited on page 22)

Zlochower, A. J. and Cohn, J. F., 1996. Vocal timing in face-to-face interaction of clinically depressed and nondepressed mothers and their 4-month-old infants. *Infant Behavior and Development*, 19, 3 (1996), 371 – 374. doi:http://dx.doi.org/10.1016/S0163-6383(96)90035-1. http://www.sciencedirect.com/science/article/pii/S0163638396900351. (cited on pages 11 and 72)