Acoustic and Conversational Speech Analysis of Depressed Adolescents and their Parents

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

Melissa Stolar
B.Sc. (Physics)
B.Eng. (Hons., Electronics and Communications)

School of Engineering
College of Science, Engineering and Health
RMIT University

August 2016
DECLARATION

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

Melissa Stolar

11/01/2017
ABSTRACT

Clinical depression is a debilitating disorder with increasing prevalence, economically a burden and linked to high suicide rates. Symptoms can show during adolescence and lead to long-term consequences in adulthood if left untreated. Current diagnosis issues are related to access, seeking treatment and subjective techniques.

An automated, discrete and efficient system, capable of detecting depression and analyzing risk factors is required. This could assist in regulating family interactions with therapies to alter patterns correlated to depression.

An audio-visual database collected by Oregon research institute (ORI-DB) provided a corpus of 68 adolescents (29 depressed and 34 non-depressed) in dyadic parent-adolescent conversations. ORI-DB has corresponding living in family environments annotations (LIFE) facilitating audio and conversation approaches.

Speech analysis in past depression studies are mostly restricted to small databases of adult speech; dependent on audio quality, gender, speaker and environment. The acoustic approach investigated a range of spectral, prosodic, cepstral, TEO and glottal features and new glottal waveform features with consistent results with past studies:

- Gender dependence increased adolescent depression detection rates
- Spectral features outperform prosodic features
- TEO are robust to noisy signals and MFCC performance degrades with noise
- Glottal waveform (G-MFCC/G-TEO) improves on speech features (TEO/MFCC)
- Combinations of subcategories and categories enhance depression detection

Important new contributions provided the following observations:

- A new roll-off range was introduced and improved with 2-stage mRMR/SVM
- Also, to our knowledge, the first use of G-MFCC in depression detection
- For the first time proposed adolescent depression detection from parent’s features
Additionally, this thesis explored depression in relation to emotional characteristics in conversation. Clinical depression is an affect (emotion) regulation disorder [24], associated with emotional disturbances [449] and has a strong correlation with expressed emotions, mood changes, and stress [21][100][16]. Furthermore, adolescent depression is strongly correlated to quality of family interaction in terms of emotions [24][134][135][136][320][144][157]. Considering the link between emotion and depression it would be expected that analysis of emotional characteristics in conversations could be applied towards detect clinical depression.

It was proposed to generate a conversation modeling system (CMS) as a complex model to capture emotional dependence patterns related to inter- and intra-influences and emotional transitions in dyadic emotional annotated sequences. The CMS is based on an Influence Model (IM) and extended to an Emotional Influence Model (EIM), Dynamic (DEIM) and Higher Order (HOEIM) with Mixed Memory Markov (HOEIM-MMM) and Kneser-Ney Smoothed ngrams (HOEIM-KN-ngram).

This thesis makes contributions to the analysis and understanding of emotional conversations in the context of depression detection and potential therapies. A first study of its kind the EIM features were analyzed, classified and can provide new insights into depression with the following major observations:

- HOEIM improved data fit, in terms of log-likelihood, as the order increased
- DEIM/HOEIM parameters generated psychologically valid interoperations of emotional influence with statistical significance between depressed and control
- Depression classification better using higher delay/order and found HOEIM had higher depression detection rates (DEIM<HOEIM-MMM<HOEIM-KN-ngram)

The major conclusions of the best overall depression detection performance in each acoustic and conversation modeling approaches as follows:

- Acoustic features: P+S (GDM-M) 98.8% and G+S (GDM-F) 97.1%
- Conversation Modeling: HOEIM-KN-ngram mRMR/SVM (GIM) 95.5%
ACKNOWLEDGEMENTS

The work presented would not have been possible without help, motivation and support I received to accomplish this effort and survive. Therefore, I would like to express my appreciation and thankfulness to all those that have helped and supported me through this tough journey.

I would like to acknowledge the Australian government for granting me the Australian Postgraduate Award (APA) and RMIT for giving me a candidate placement.

Primarily, I would like to express gratitude to my supervisor Dr Margaret Lech for the opportunity, encouragement and for convincing me to do a Ph.D. I would like to thank you for the constant support, enthusiasm, and guidance throughout my time at university. Also I appreciated all your opinions, ideas and knowledge and would not have been able to finish without you.

Additionally, I would like to thank my secondary supervisors Prof. Ian Burnett and Dr. Eva Cheng for their insights and suggestions.

My sincere thanks also goes to Professor Nicholas Allen from the University of Oregon for his time, valuable insights in regards to the psychological aspects of this thesis. Thanks for the effort to improve the quality of our publications by expanding on psychological concepts. I would also like to thank Dr Lisa Sheeber, from the Oregon Research Institute, for providing a database and expertise. Without data or the collaborative efforts between psychology and engineering fields the research conducted would not have been possible.
I would also like to thank fellow students and DSP lab colleagues at RMIT for discussions, whether it is amusing distractions or actually helpful work related conversations. In particular, I am grateful to Haytham Fayek for the comedic relief, advice and most importantly helping with a title decision ☺.

Finally, I want to thank all my family for the time you spent to listen to my problems and try to help during my time as a student. I appreciate everything you have done for all the encouragement, support and belief you had in me to pursue my dreams.

Thank you to my parents, Joanna and John, for giving me life, putting up with me and keeping me alive to this point. I am grateful for all the time spent with my grandparents, their unwavering support, friendship and their unconditional love but “I love you the most”.

A big thank you to my younger sister, Shannon, for all her help proofreading and editing my work. One day when you write your thesis its pay back time. For the first and only time ever I am thankful for my twin sister, Bianca, for your lack of encouragement and support and the direct competition giving me constant motivation so now I can say “I win”! ☺.
I dedicate this thesis to my grandfather,

Boris Stolar
LIST OF PUBLICATIONS

Journal Publications:


Conference publications:


# TABLE OF CONTENTS

DECLARATION ........................................................................................................... ii
ABSTRACT .................................................................................................................. iii
ACKNOWLEDGEMENTS .............................................................................................. v
DEDICATION ............................................................................................................... vii
LIST OF PUBLICATIONS ............................................................................................ viii
TABLE OF CONTENTS .............................................................................................. ix
LIST OF TABLES ......................................................................................................... xv
LIST OF ABBREVIATIONS ......................................................................................... xxxi

## CHAPTER ONE:

1.1 Preview ................................................................................................................... 1
1.2 Background and Problem Statement/Description ................................................. 1
1.3 Thesis Aims and Research Questions ..................................................................... 8
1.4 Thesis Scope and Limitations ................................................................................ 11
1.5 Thesis Contributions ............................................................................................. 13
1.6 Thesis Structure .................................................................................................... 17

## CHAPTER TWO:

2.1 Preview .................................................................................................................. 19
2.2 Literature Review on Emotion Recognition ......................................................... 19
   2.2.1 Background of Emotion Recognition ............................................................ 20
   2.2.2 Overview of speech production in relation to signal processing .................... 21
   2.2.3 Physiological rationale for acoustic AER from speech production ............... 23
   2.2.4 Existing Automatic Emotion recognition studies using acoustic features .......... 24
   2.2.5 Deep Machine Learning and classification .................................................... 31
      2.2.5.1 Deep Neural Network (DNN) Applications ............................................. 31
      2.2.5.2 Deep Learning for emotion recognition .................................................. 32
      2.2.5.3 Multichannel Classification .................................................................... 33
   2.2.6 Summary Emotion Recognition Literature .................................................. 34
2.3 Literature on Depression Recognition from Speech ............................................. 35
   2.3.1 Background of Automatic Depression Detection Modes ................................ 35
   2.3.2 Physiological Correlation of Vocal Indicators of Depression ......................... 37
   2.3.3 Vocal Indicators of depression ....................................................................... 38
   2.3.4 Acoustic speech parameters in machine learning Depression Detection .......... 41
   2.3.5 Limitation with acoustic depression recognition studies .................................. 45
   2.3.6 Summary Speech Depression Recognition Literature .................................... 47
2.4 Literature Review for Depression Recognition using Conversation Analysis ........ 48
   2.4.1 Psychological validity of depression diagnosis from conversations ................ 48
   2.4.2 Studies using psychological models of conversations for depression ............... 50
   2.4.3 Studies using generative and probability models for social analysis ............... 51
   2.4.4 Summary Conversation Analysis Depression Recognition Literature ................ 54
2.5 Literature for Emotion Time Series Forecasting .................................................... 55
   2.5.1 Methodologies and studies related to time forecasting ..................................... 55
   2.5.2 Time series forecasting relation to affective forecasting .................................. 57
   2.5.3 Summary of Emotion Forecasting Literature ................................................. 58
2.6 Summary ................................................................................................................ 59
CHAPTER THREE: .................................................................................................................. 60
3.1 Preview .......................................................................................................................... 60
3.2 Depression Database ................................................................................................. 60
3.3 Oregon Research Institute (ORI) Database ............................................................... 62
   3.3.1 Participants ............................................................................................................. 62
   3.3.2 Recording Setup .................................................................................................. 62
   3.3.3 Conversational data collection procedure ............................................................ 63
   3.3.4 Database Annotation Living-In-Family Environments Coding System ............... 64
   3.3.5 Speech Corpus – Experimental Group (Depressed and Control) ....................... 65
3.4 Emotion Database .........................................................................................................66
3.5 Berlin Emotional Speech Database (EMO-DB) ............................................................ 67
   3.5.1 Participants ............................................................................................................. 67
   3.5.2 Recording Setup .................................................................................................. 67
   3.5.3 Emotional utterance collection procedure ............................................................ 67
   3.5.4 Speech Corpus – Experimental Group (Emotions) ............................................ 67

CHAPTER FOUR: .................................................................................................................. 68
4.1 Preview .......................................................................................................................... 68
4.2 Automatic Emotion Recognition ................................................................................ 69
4.3 Pre-Processing .............................................................................................................. 70
   4.3.1 Voiced Activity Detector (VAD) ......................................................................... 70
   4.3.2 Glottal Waveform Estimation ............................................................................ 72
4.4 Feature Extraction ........................................................................................................ 74
   4.4.1 Cepstral-Mel frequency cepstral coefficients (MFCC) ........................................ 74
   4.4.2 Teager energy operators (TEO) ........................................................................... 76
      4.4.2.1 Area under TEO critical band autocorrelation envelopes (TEO-CB-Auto-Env) ......................................................... 77
   4.4.3 Glottal Waveform Features .............................................................................. 79
   4.4.4 Delta (Δ) and Delta-Delta (Δ-Δ) Coefficients ...................................................... 79
4.5 2D image generation .................................................................................................... 80
4.6 Modeling and Classification systems .......................................................................... 80
   4.6.1 Gaussian Mixture Model (GMM) ....................................................................... 81
   4.6.2 Support Vector Machines (SVM) ....................................................................... 83
   4.6.3 Neural Network .................................................................................................... 89
   4.6.4 Deep Neural Network (DNN) ............................................................................ 91
4.7 New Optimized Multichannel Weighted Deep Neural Network (OMC-DNN) ........ 96
   4.7.1 Optimized Multichannel DNN (OMC-DNN) system description ....................... 97
   4.7.2 Determining the Optimal weight ....................................................................... 99
   4.7.3 Optimization algorithm .....................................................................................100
      4.7.3.1 Simulated Annealing ..................................................................................101
      4.7.3.2 Hybrid: Simulated Annealing with Genetic Algorithm (SA/GA) ..............103
4.8 Acoustic Emotion Recognition Modeling and Classification Setup .........................104
   4.8.1 Single channel classifier setups ..........................................................................104
   4.8.2 Multichannel channel classifiers setups ..............................................................106
4.9 Experimental setups ...................................................................................................107
4.10 Evaluation methods ....................................................................................................108
4.11 Single Channel Emotion Recognition Results ...........................................................109
   4.11.1 Binary AER using SC-DNN with MFCC and TEO features (EXP1)..............109
4.11.2 Binary AER using SC-DNN with G-MFCC and G-TEO features (EXP2) ................................................................. 109
4.11.3 Comparison of Glottal and Speech Features using DNN for Binary AER .................................................................................................................. 110
4.11.4 Effectiveness of GDMs and GIM for AER using SC-DNN (EXP3) ........... 111
4.11.5 GDMs and GIM for Multiclass AER using SC-DNN (EXP4) ................... 114
4.11.6 Comparison of emotion recognition with benchmark classifiers (EXP5) .................................................................................................................. 116
4.12 Optimized Multichannel DNN (OMC-DNN) emotion recognition results .... 118
4.12.1 Effectiveness of emotion recognition with OMC-DNN (EXP6) ............... 119
  4.12.1.1 Sets A-D as 3-channel OMC-DNN .................................................. 119
  4.12.1.2 Sets E-G as 4 channel OMC-DNN ................................................... 122
4.13 Discussion and Summary ..................................................................... 124

CHAPTER FIVE: ................................................................................. 129
5.1 Preview ....................................................................................... 129
5.2 Oregon Research Institute (ORI) annotations corpus ............................... 130
  5.2.1 ORI-DB Annotation Corpus– Statistical Analysis ................................ 131
5.3 Influence Model background knowledge ................................................. 134
  5.3.1 Introduction to the Influence Model .................................................... 135
  5.3.2 Learning the influence Model (Optimization Algorithm) ...................... 136
  5.3.3 Introducing emotional interactions to Influence Models (EIM) ............. 139
  5.3.4 Introducing a Dynamic Emotional Influence Model (DEIM) ............... 141
  5.3.5 Introducing a Higher Order Emotional Influence Model (HOEIM) ....... 142
     5.3.5.1 Higher Order EIM using Mixed Memory Markov Model (HOEIM-MMM) ................................................................................................. 144
     5.3.5.2 Higher Order EIM using n-gram method (HOEIM-ngram) .......... 145
     5.3.5.3 HOEIM Modified Kneser-Ney smoothed ngram (HOEIM-KN-ngram) ............................................................................................................. 146
5.4 Depression Detection using Influence Models ........................................ 148
  5.4.1 Feature Extraction from the Influence Models ...................................... 148
  5.4.2 Modeling and Classification Setup .................................................... 149
     5.4.2.1 Support Vector Machine (SVM) ....................................................... 149
     5.4.2.2 Neural Network (NN) ..................................................................... 150
     5.4.2.3 Optimized Neural Network (ο-NN) .................................................. 150
5.5 Experimental Setup .......................................................................... 152
5.6 Evaluation Methods .......................................................................... 154
5.7 Experimental results on the Influence Model generation ........................ 155
  5.7.1 Preliminary experiment using Auto and Cross Correlation (EXPprelim) .............................................................................................................. 155
  5.7.2 Compare optimization methods of EIM (EXP1) .................................... 158
  5.7.3 Determine effect of delay using Dynamic Influence Model (EXP2) ...... 158
  5.7.4 Determination of higher order dependence of EIM using HOEIM (EXP3) ........................................................................................................... 159
5.8 Experimental Results of Influence model Parameters Analysis and
  Statistical ......................................................................................... 162
  5.8.1 Emotional Transition Probabilities Statistical Testing (EXP4) ............... 162
  5.8.2 Influence Coefficients for each EIM variation (EXP5) ......................... 165
     5.8.2.1 DEIM Influence Coefficients and Discussion ................................... 165
     5.8.2.2 DEIM Influence Coefficients Statistical Tests ................................. 172
5.8.2.3 HOEIM Influence Coefficients and Discussion ................................173
5.8.2.4 HOEIM Influence Coefficients Statistical tests ..............................178
5.9 Experimental Results of Depression Classification ...............................179
  5.9.1 Depression detection using the DEIM with SVM (EXP6) ..................179
  5.9.2 Effectiveness of DEIM for depression classification using NN (EXP7) .......180
  5.9.3 Optimized NN for Depression Classification using the DEIM (EXP8) ..........181
  5.9.4 Compare the DEIM classification system for depression detection ..........183
  5.9.5 Effectiveness of HOEIM parameters for depression classification (EXP9) .................................................................185
  5.9.5.1 Depression detection using HOEIM-MMM parameters ....................185
  5.9.5.2 Depression detection using HOEIM-KN-ngram parameters .................186
  5.9.6 Depression detection using optimized EIM parameters (EXP10) ..........187
  5.9.7 Comparison of HOEIM depression classification performance ..........189
5.10 Discussion and Summary ..................................................................191

CHAPTER SIX: .........................................................................................195
6.1 Preview ..............................................................................................195
6.2 Methodology ......................................................................................195
  6.2.1 Introduction of the Random Walk ..................................................195
  6.2.1.1 Biased Random Walk ..............................................................196
  6.2.1.2 Extension to higher order random walk (HORW) .......................197
  6.2.1.3 New interacting random walk (IRW) ........................................197
  6.2.1.4 2D representation of the random walk methods .........................197
  6.2.2 Non Linear Autoregression (NAR) models for prediction .................200
    6.2.2.1 NAR with Neural Networks (NAR-NN and NARX-NN) ...............201
    6.2.2.2 NAR with Adaptive Neuro-Fuzzy Inference System (ANFIS) ........204
6.3 Experimental Setups ...........................................................................206
  6.3.1 Experimental Setup Random Walks ...............................................206
  6.3.2 Experimental Setup NN and ANFIS methods of NAR(X) models ........207
6.4 Experimental results ..........................................................................209
  6.4.1 Experimental results implemented with RW ....................................209
    6.4.1.1 Effectiveness of higher order random walks (HORW) (EXP1) ....209
    6.4.1.2 Effectiveness of the interacting random walk (IRW) (EXP2) ......209
    6.4.1.3 Investigate 2D RW for link between depressed interaction (EXP3) ..210
  6.4.2 Experimental results implemented with NAR models ....................214
    6.4.2.1 Prediction using Speaker and Class Independent (SI-CI) NAR-NN (EXP1) .................................................................214
    6.4.2.2 Effect of external factor (NARX-NN) on prediction (SI-CI) (EXP2) .214
    6.4.2.3 Determine NAR(X)-NN depression dependence (SI-CD) (EXP3) ....215
    6.4.2.4 Compare NAR(X)-NN with Speaker Dependence (SD-CI) (EXP4) ...216
    6.4.2.5 Compare NAR(X)-NN with Speaker and Class Dependence (SD-CD) (EXP5) .................................................................218
    6.4.2.6 Results comparing ANFIS with external output and without (EXP6) .219
    6.4.2.7 Comparison of NAR prediction implemented with ANFIS and NN (EXP7) ......................................................................220
6.5 Summary and Conclusions ..................................................................220
CHAPTER SEVEN: ................................................................. 222
7.1 Preview .............................................................................. 222
7.2 Pre-processing................................................................. 223
  7.2.1 Decimate Sampling rate .................................................. 223
  7.2.2 Background noise removal ............................................. 224
  7.2.3 Source Separation .......................................................... 225
  7.2.4 Voiced Speech Windows ................................................ 227
7.3 Acoustic Feature Extraction .............................................. 228
  7.3.1 Prosodic ....................................................................... 230
    7.3.1.1 Fundamental Frequency (F0) .................................... 230
    7.3.1.2 Log energy (LogE) .................................................. 231
    7.3.1.3 Jitter ..................................................................... 231
    7.3.1.4 Shimmer ............................................................... 231
  7.3.2 Spectral ........................................................................ 232
    7.3.2.1 Formants ............................................................... 232
    7.3.2.2 Power spectral density ............................................ 233
    7.3.2.3 Spectral flux .......................................................... 233
    7.3.2.4 Spectral centroid .................................................... 233
    7.3.2.5 Spectral entropy ..................................................... 234
    7.3.2.6 Spectral roll-off ..................................................... 235
  7.3.3 Cepstral (MFCC) .......................................................... 236
    7.3.4 TEO critical-band autocorrelation envelope (TEO-CB-Auto-Env) ........................................ 236
  7.3.5 Glottal Features ............................................................ 237
    7.3.5.1 Glottal Time Domain (GTD) .................................... 238
    7.3.5.2 Glottal Frequency Domain (GFD) ............................. 240
    7.3.5.3 Glottal MFCC and Glottal TEO (G-MFCC and G-TEO) ..................................................... 240
7.4 Modeling and Classification Setup ...................................... 241
7.5 Experimental setup for acoustic classification of depression ................................................................. 242
7.6 Evaluation Methods .......................................................... 245
7.7 Experimental results .......................................................... 246
  7.7.1 Statistical Results- MANOVA and ANOVA (EXP1) .............. 246
  7.7.2 Effectiveness of Gender Dependent vs. Gender Independent Models (EXP2) .......................................................... 247
  7.7.3 Optimize Feature Selection from Spectral Roll-off Coefficients (EXP3) .......................................................... 248
  7.7.4 Effectiveness of Spectral and Prosodic Feature Sub-Categories (EXP4) .............................................. 252
  7.7.5 Effectiveness of Prosodic and Spectral Feature Categories (EXP5) .............................................. 254
  7.7.6 Cepstral (MFCC) and TEO-based (TEO-CB-Auto-Env) Features (EXP6) .................................................. 256
  7.7.7 Robustness of MFCC and TEO-CB-Auto-Env with AWGN (EXP7) ........................................... 258
  7.7.8 Glottal Category (G) time (GTD) and frequency (GFD) features (EXP8) .............................................. 260
  7.7.9 Glottal Waveform (GW) features (G-MFCC and G-TEO) (EXP9) .............................................. 261
  7.7.10 Combine glottal category (G) with categories (S, P, C, TEO, GW) (EXP10) ........................................... 262
  7.7.11 Combine Spectral category (S) with categories (P, C, TEO, G, GW) (EXP11) ........................................... 265
  7.7.12 Depression detection in adolescents using parents speech (EXP12) ........................................... 267
7.8 Discussion and Conclusions .................................................. 270
CHAPTER EIGHT: ........................................................................................................276
8.1 Preview .................................................................................................................276
8.2 Discussion and Summary .......................................................................................276
  8.2.1 Emotion Related Studies ...............................................................................277
    8.2.1.1 Automatic Emotion Recognition (AER) Studies .......................................277
    8.2.1.2 Affect Prediction Studies .......................................................................279
  8.2.2 Depression Related Studies ..........................................................................281
    8.2.2.1 Conversation Modeling System (CMS) Approach ....................................281
    8.2.2.2 Acoustic Speech Feature modeling Approach ...........................................283
  8.2.3 Acoustic Features versus Conversation Modeling (EIM) Approaches ..........286
  8.2.4 Performance overview of depression classification compared to past
    literature ...............................................................................................................289
  8.3 Future Work ........................................................................................................291

REFERENCES ............................................................................................................293
LIST OF TABLES

Table 3-1 GENDER DISTRIBUTION OF DEPRESSED AND NON-DEPRESSED ADOLESCENTS AND THEIR RESPECTIVE PARENT IN ORI-DB DYADIC CONVERSATIONS.................................................................65

Table 3-2 DISTRIBUTION OF EMOTION IN THE BERLIN EMOTIONAL SPEECH DATABASE (EMO-DB) FOR BOTH MALE AND FEMALE SUBJECTS ......................67

Table 4-1 TEO SUB-BAND’S UPPER (U), LOWER (L) FREQUENCIES (HZ) AND BANDWIDTH (BW) DERIVED FROM THE WAVELET PACKETS .........................77

Table 4-2 BINARY OMC-DNN DECISION MATRIX TO CONVERT THE WEIGHTED RESULT, sx1, TO DETERMINE THE FINAL RESULT, hxi, AND THE CORRESPONDING PREDICTED CLASS (EMOTION OR NEUTRAL)..........................98

Table 4-3 MULTICLASS OMC-DNN DECISION MATRIX TO CONVERT THE WEIGHTED RESULT, sx1, TO DETERMINE THE FINAL RESULT, hxi, AND THE CORRESPONDING PREDICTED CLASS..................................................98

Table 4-4 AVERAGE PERFORMANCE OF BINARY EMOTION RECOGNITION USING THE SINGLE CHANNEL DNN APPROACH COMPARING MFCC AND TEO FEATURES USING THE EMO-DB (A: ANGER, B: BORED, D: DISGUST, F: FEAR, J: JOY AND S: SAD)................................................................................................................109

Table 4-5 AVERAGE PERFORMANCE OF BINARY AER USING SC-DNN COMPARING GLOTTAL WAVEFORM FEATURES (G-MFCC AND G-TEO) USING THE EMO-DB ...........................................................................................................109

Table 4-6 AVERAGE ACCURACY, SENSITIVITY AND SPECIFICITY OF BINARY AER USING A SC-DNN COMPARING MFCC AND TEO FEATURES WITH BOTH GENDER DEPENDENT MODELS (GDM-M AND GDM-F)..........................111

Table 4-7 AVERAGE ACCURACY, SENSITIVITY AND SPECIFICITY OF BINARY AER USING A SC-DNN COMPARING G-MFCC AND G-TEO FEATURES WITH BOTH GENDER DEPENDENT MODELS (GDM-M AND GDM-F)..........................111

Table 4-8 COMPARING AVERAGE PERFORMANCE OF GDM (MALE AND FEMALE) AND GIM OF THE 6 BINARY AER SETUPS (ANGER, BORED, DISGUST, FEAR, JOY AND SAD) USING SC-DNNs WITH S-MFCC, G-MFCC, S-TEO, G-TEO ....113

Table 4-9 CONFUSION MATRIX FOR GIM MULTICLASS AER WITH SC-DNN USING S-MFCC AND S-TEO FEATURES (A: ANGER, B: BORED, D: DISGUST, F: FEAR, J: JOY AND S: SAD)........................................................................................................114

Table 4-10 CONFUSION MATRIX FOR GIM MULTICLASS AER WITH SC-DNN USING G-MFCC AND G-TEO FEATURES (A: ANGER, B: BORED, D: DISGUST, F: FEAR, J: JOY AND S: SAD)......................................................114
Table 4-11 AVERAGE ACCURACY (%) OF MULTICLASS (7-CLASSES) SC-DNN USING THE S-MFCC, S-TEO, G-MFCC AND G-TEO FEATURES WITH GENDER DEPENDENCE (GDM-M AND GDM-F) AND GENDER INDEPENDENCE (GIM) ....115

Table 4-12 AVERAGE ACCURACY (%) OF 6 BINARY AER SYSTEMS WITH A COMPARISON OF DNN AND BENCHMARK CLASSIFIERS (NN, GMM AND SVM) USING S-MFCC, G-MFCC, S-TEO, G-TEO WITH 2D OR 1D FEATURES FOR GDM AND GIM ..........................................................116

Table 4-13 P-VALUES OF PAIRED T-TESTS STATISTICALLY ANALYZING BINARY DNN AER ACCURACIES COMPARED TO BENCHMARK CLASSIFIERS (NN, GMM, SVM) FOR EACH FEATURE SET WITH STATISTICAL SIGNIFICANCE FOR P>0.05 .................................................................................................................117

Table 4-14 AVERAGE ACCURACY FOR MULTICLASS AER USING 2D FEATURES IN THE DNN AND FEATURE VECTORS FOR THE NN AND GMM BASELINE CLASSIFIERS. RESULTS ARE SHOWN FOR THE GIM AND GDMs AND FOUR FEATURE SETS ........................................................................................................117

Table 4-15 OUTLINE OF THE FOUR DIFFERENT 3-CHANNEL FEATURE SETS (A-D) FOR THE OMC-DNN EMOTION RECOGNITION EXPERIMENTS ..............................................119

Table 4-16 AVERAGE WEIGHT ALLOCATION, OF SIX BINARY AER PAIRS, FOR FEATURE SETS A, B, C AND D (STATIC, DERIVATIVE, DOUBLE DERIVATIVE FEATURE CHANNELS) OF THE 3-CHANNEL OMC-DNN IN GIM AND GDM ......119

Table 4-17 AVERAGE ACCURACY (%) FOR THE SIX BINARY AER TESTS USING THE 3-CHANNEL OMC-DNN WITH FOUR DIFFERENT FEATURE SETS (A, B, C AND D) WITH GIM AND GDM (MALE AND FEMALE) ......................................................120

Table 4-18 AVERAGE TEST ACCURACY (%), ACROSS 6 BINARY PAIRS, OF THE THREE INDIVIDUAL CHANNELS AND THE OVERALL 3-CHANNEL OMC-DNN FOR FEATURE SETS A-D WITH GIM AND GDM ..............................................120

Table 4-19 AVERAGE WEIGHT ALLOCATION, FOR FEATURE SETS A, B, C AND D (STATIC, DERIVATIVE, DOUBLE DERIVATIVE FEATURE CHANNELS), OF THE 3-CHANNEL OMC-DNN IN GIM MULTICLASS (7-CLASS) AER ........................................121

Table 4-20 AVERAGE ACCURACY (%) FOR MULTICLASS AER OF INDIVIDUAL SC-DNN AND THE ENTIRE OMC-DNN USING A 3-CHANNEL SETUP WITH FOUR DIFFERENT FEATURE SETS (A, B, C AND D) IN A GENDER INDEPENDENT MODEL (GIM) ........................................................................................................121

Table 4-21 OUTLINE OF THE THREE DIFFERENT 4-CHANNEL FEATURE SETS (E-G) FOR THE OMC-DNN EMOTION RECOGNITION EXPERIMENTS ...........................................122

Table 4-22 AVERAGE WEIGHT ALLOCATION, OF SIX BINARY AER PAIRS, FOR FEATURE SETS E, F AND G (S-TEO, S-MFCC, G-TEO AND G-MFCC FEATURE CHANNELS) OF THE 4-CHANNEL OMC-DNN IN GIM AND GDM ......122
Table 4-23 AVERAGE ACCURACY (%) FOR THE BINARY AER TESTS USING THE 4-CHANNEL OMC-DNN WITH THREE DIFFERENT FEATURE SETS E-G FOR BOTH GIM AND GDM ................................................................. 123

Table 4-24 AVERAGE TEST ACCURACY (%), ACROSS 6 BINARY PAIRS, OF EACH INDIVIDUAL CHANNEL AND THE OVERALL 4 CHANNEL OMC-DNN FOR FEATURE SETS E-G WITH GIM AND GDM ................................................................. 123

Table 4-25 SUMMARY OF THE BINARY OMC-DNN RESULTS, FOR EVERY FEATURE SET (A, B, C, D, E, F), GIVING THE AVERAGE TEST ACCURACIES AND THE CORRESPONDING TOP INDIVIDUAL FEATURE CHANNEL OF THE SC-DNNs. ........ 125

Table 4-26 SUMMARY OF BINARY OMC-DNN RESULTS WITH THE MINIMUM AND MAXIMUM ACCURACY OF EACH TYPE OF FEATURE SETS INCLUDING 3-CHANNEL (A-D) AND 4-CHANNEL (E-G) ................................................................. 126

Table 4-27 SUMMARY OF RELATED MULTICLASS EMOTION RECOGNITION PUBLICATIONS OUTLINING THE ACOUSTIC FEATURES, CLASSIFICATION SYSTEM, DATABASE AND ACCURACY (%). ................................................................. 127

Table 5-1 COMPARISON OF OBJECTIVE FUNCTION (LOG-LIKELIHOOD (LL) OR MEAN SQUARE ERROR (MSE)) AND BOTH CONSTRAINTS (TYPE 1 AND TYPE 2) FOR THE SIX LEARNING METHODS: GD-GRADIENT DESCENT, EM-EXPECTATION MAXIMIZATION, MVR-MULTIVARIATE LINEAR REGRESSION, QN-QUASI-NEWTON, SA-SIMULATED ANNEALING, SA/GA-COMBINED SIMULATED ANNEALING AND GENETIC ALGORITHM ................................................................. 138

Table 5-2 NUMBER OF INFLUENCE MODEL PARAMETERS (INFLUENCE COEFFICIENTS (IC) AND TRANSITION PROBABILITIES (P)) IN EACH APPROACH, WHERE M IS THE NUMBER OF SPEAKERS, Q IS THE NUMBER OF STATES AND N IS THE ORDER ......................................................................................................................... 149

Table 5-3 STATISTICAL DISTRIBUTION OF THE TIME-DELAY (Nq=1, ..., 7) FOR THE MAXIMUM MMM WEIGHT OCCURRENCES INCLUDING THE MEDIAN, STANDARD DEVIATION AND AVERAGE TIME-DELAY OF EACH WEIGHT (ω11, ω12, ω21, ω22) ................................................................................................................................. 161

Table 5-4 T-TEST OF CHILDREN’S SELF-TRANSITION PROBABILITIES (P11) BETWEEN DEPRESSED AND CONTROL. SIGNIFICANT TRANSITIONS, p<0.05, ARE IN BOLD AND BLUE SHADE DESIGNATE TRANSITIONS MORE LIKELY FROM DEPRESSED ........................................................................................................... 162

Table 5-5 T-TEST OF CHILDREN’S CROSS-TRANSITION PROBABILITIES (P12) BETWEEN DEPRESSED AND CONTROL. SIGNIFICANT TRANSITIONS, p<0.05, ARE IN BOLD AND BLUE SHADE DESIGNATE TRANSITIONS MORE LIKELY FROM DEPRESSED ........................................................................................................... 163

Table 5-6 T-TEST OF PARENT’S CROSS-TRANSITION PROBABILITIES (P21) BETWEEN DEPRESSED AND CONTROL ADOLESCENTS. SIGNIFICANT TRANSITIONS, p<0.05, ARE IN BOLD AND BLUE SHADE DESIGNATE TRANSITIONS MORE LIKELY FROM PARENTS OF DEPRESSED CHILDREN ........ 163
Table 5-7 T-test of Parent's Self-Transition Probabilities (P_{22}) between Depressed and Control. Significant Transitions, p<0.05, are in bold and blue shad designate transitions more likely from Parents of Depressed Children

Table 5-8 Statistical Analysis of the Entire Set of DEIM Generated Influence Coefficients (θ_{11}, θ_{12}, θ_{21}, θ_{22}) Comparing Between Depressed and Non-Depressed Adolescents. Results give the MANOVA Wilks Lambda p-value, where Significant (p<0.05) parameters are shaded, for each topic (EPI, FCI, PSI and ALL TOPICS) and DEIM Delay (t-1, ..., t-10).

Table 5-9 Statistical Analysis of DEIM Generated Influence Coefficients (θ_{11}, θ_{12}, θ_{21}, θ_{22}) Comparing Depressed and Non-Depressed Adolescents. Results illustrate the T-test p-values, where shaded cells denote parameters that are significant (p<0.05), for each topic (EPI, FCI, PSI) and DEIM Delay (t-1, ..., t-10).

Table 5-10 Statistical Analysis of the Entire Set of HOEIM Generated Influence Coefficients (θ_{11}, θ_{12}, θ_{21}, θ_{22}) Comparing Between Depressed and Non-Depressed Subjects with MANOVA Results given for each Order (N) and Interaction as the Wilks Lambda Value (p<0.05 significance).

Table 5-11 Depression Classification Performance using a SVM and the Entire DEIM Feature Set (P_{11}, P_{12}, P_{21}, P_{22}, θ_{11}, θ_{12}, θ_{21}, θ_{22}) for each Delay and Interaction.

Table 5-12 Depression Classification Performance using a NN (25 nodes) and the Entire DEIM Feature Set (P_{11}, P_{12}, P_{21}, P_{22}, θ_{11}, θ_{12}, θ_{21}, θ_{22}) for each Delay and Interaction.

Table 5-13 Optimum Number of Hidden Nodes in a NN for the Best Performance in Depression Classification using 5-fold CV given for each DEIM Feature Set and for each Interaction (EPI, FCI, PSI).

Table 5-14 Depression Classification Performance using the Optimized Hidden Node Neural Network (O-NN) with the Entire DEIM Feature Set (P_{11}, P_{12}, P_{21}, P_{22}, θ_{11}, θ_{12}, θ_{21}, θ_{22}) for each Delay and Interaction.

Table 5-15 Comparison of Classification Methods (O-NN, NN, SVM) using the Entire Set of DEIM Parameters (P_{11}, P_{12}, P_{21}, P_{22}, θ_{11}, θ_{12}, θ_{21}, θ_{22}) as a Feature Set. Results are given as an average accuracy (%) of all ten DEIM Delays (t-1, t-2, ..., t-10).

Table 5-16 Depression Classification Performance using SVM with the Entire Set of HOEIM-MMM Features (P_{11}, P_{12}, P_{21}, P_{22}, θ_{11}, θ_{12}, θ_{21}, θ_{22}) for each Order (N) and Interaction.
Table 5-17 Depression Classification Performance using SVM with the entire set of KN-NGram HOEIM Features (P_{11}, P_{12}, P_{21}, P_{22}, \theta_{11}, \theta_{12}, \theta_{21}, \theta_{22}) for each Order (N) and Interaction ... 186

Table 5-18 Depression Classification Performance using SVM and mRMR in 2-Stage Feature Selection Approach of HOEIM-KN-NGram-Parameters (P_{11}, P_{12}, P_{21}, P_{22}, \theta_{11}, \theta_{12}, \theta_{21}, \theta_{22}) for each Order (N) and Interaction ... 188

Table 5-19 Summary of SVM Depression Classification Accuracy, Averaged Across HOEIM Orders, Comparing HOEIM Feature Extraction Methods of MMM and KN-NGram including with a 2-stage mRMR/SVM Optimization ............................................................... 190

Table 5-20 Ranking of Depression Classification Average Accuracy (%) of each Interaction (EPI, FCI, PSI) using the ORI-DB LIFE annotations construct states (positive, negative, neutral, silence) for the Influence Model Approach. The cells shaded darkest to lightest denote the: EIM, DEIM, HOEIM-MMM, HOEIM-KN-NGram and HOEIM-KN-NGram using mRMR feature selection ............................................................... 194

Table 6-1 Comparison Average Relative Accuracy (%) of Affect Prediction using Higher Order Simple Random Walk (HORW) with 10, 25, 50 and 250 steps. Shaded Cells Indicate Setups with Higher Accuracy Compared to the IRW (Table 6-1) ............................................................... 209

Table 6-2 Comparison Average Relative Accuracy (%) of Affect Prediction using Higher Order Interacting Random Walk (IRW) with 10, 25, 50 and 250 steps. Shaded Cells Indicate Setups with Higher Accuracy Compared to the HORW (Table 6-1) ............................................................... 209

Table 6-3 Average Accuracy (%) of Correct Final Quadrant Position Comparing the Interacting Random Walk (IRW) and Random Walk (RW) for two Different Iteration Lengths (25 and 250 steps) and Orders 1-5 .................................................................................. 211

Table 6-4 Average Accuracy (%) of Affect Prediction using the Speaker Independent and Class Independent (SI-CI) NAR-NN with a Comparison of Orders for each Interaction (EPI, FCI, PSI). Bold Cell Indicates the Interaction and Order with the Highest Accuracy for the Given Setup ............................................................................. 214

Table 6-5 Average Accuracy (%) of Affect Prediction using the Speaker Independent and Class Independent (SI-CI) NARX-NN with a Comparison of Orders for each Interaction (EPI, FCI, PSI) ............................................................................. 214

Table 6-6 Comparison of NAR-NN and NARX-NN Average Affect Prediction Accuracy (%) using Speaker Independent-Class Dependent Model for Depressed (SI-CD-D) (Left) and Non-Depressed (SI-CD-ND) (Right) Models ............................................................................. 215

xix
Table 6-7 COMPARISON OF NAR-NN AND NARX-NN AVERAGE AFFECT PREDICTION ACCURACY (%) FOR CLASS INDEPENDENT AND SPEAKER DEPENDENT MODEL (SD-Cl) ................................................................. 217

Table 6-8 COMPARISON OF NAR-NN AND NARX-NN AVERAGE AFFECT PREDICTION ACCURACY (%) USING SPEAKER DEPENDENT-CLASS DEPENDENCE MODEL WITH 29 DEPRESSED CONVERSATIONS (SD-CD-D) (LEFT) AND 34 NON-DEPRESSED CONVERSATIONS (SD-CD-ND) (RIGHT) ........................................... 218

Table 6-9 Comparison of NAR-ANFIS and NARX-ANFIS Average Affect Prediction Accuracy (%) using Speaker Dependent-Class Dependent Model with 29 Depressed Subject (SD-CD-D (Left)) and 34 Non-Depressed Subject (SD-CD-ND (Right)) .................................................................................................................. 219

Table 7-1 SUMMARY OF ACOUSTIC FEATURE CATEGORIES (PROSODIC (P), SPECTRAL (S), CEPstral (C), TEager-BASED (T), GLOTTal (G) AND GLOTTAL WAVEFORM (GW)) AND THE INCORPORATED SUB-CATEGORY FEATURE SETS IN ADDITION TO THE CORRESPONDING NUMBER OF FEATURE COEFFICIENTS .............................................................................................................. 229

Table 7-2 MANOVA AND ANOVA ANALYSIS COMPARING DEPRESSED AND CONTROL SUBCATEGORY FEATURES FOR MALE AND FEMALE ADOLESCENTS. WHERE "+/

---

Table 7-3 ACCURACY (%), SENSITIVITY AND SPECIFICITY OF SVM ADOLESCENT DEPRESSION DETECTION USING THE FUNDAMENTAL FREQUENCY FEATURE (F0) OF ADOLESCENTS COMPARING GIM AND GDMS FOR EACH INTERACTION (EPI, FCI, PSI) .......................................................................................................................... 247

Table 7-4 ANOVA ANALYSIS ON DEPRESSED AND CONTROL SPECTRAL ROLL-OFF FEATURES FOR MALE AND FEMALE ADOLESCENTS WHERE "<" DENOTES

---

Table 7-5 DEPRESSION CLASSIFICATION ACCURACY (%), SENSITIVITY AND SPECIFICITY USING A 2-STAGE FILTER-WRAPPER FEATURE SELECTION (mRMR/SVM) FOR THE SPECTRAL ROLL-OFF PARAMETERS COMPARED TO THE ENTIRE FEATURE SET ........................................................................................................... 251

Table 7-6 DEPRESSION CLASSIFICATION ACCURACY, SENSITIVITY AND SPECIFICITY USING SVM WITH PROSODIC (P) AND SPECTRAL (S) FEATURES FOR MALE SUBJECTS (GDM-M) FOR EACH INTERACTION (EPI, FCI, PSI) ........................................ 252

Table 7-7 DEPRESSION CLASSIFICATION ACCURACY, SENSITIVITY AND SPECIFICITY USING SVM WITH PROSODIC (P) AND SPECTRAL (S) FEATURES FOR FEMALE SUBJECTS (GDM-F) FOR EACH INTERACTION (EPI, FCI, PSI) ........................................... 252

Table 7-8 DEPRESSION CLASSIFICATION ACCURACY, SENSITIVITY AND SPECIFICITY USING SVM WITH CATEGORICAL PROSODIC, SPECTRAL* AND SPECTRAL (ROLL-OFF) FEATURE SETS FOR GDM-M AND GDM-F AND EACH INTERACTION ........................................................................................................ 254

XX
Table 7-9 **Accuracy Difference of P, S*, and S Categories Compared to the Maximum, Minimum and Average of the Individual Sub-Categorical Features** ......................................................... 255

Table 7-10 **Depression Classification Accuracy (%)**, Sensitivity and Specificity Using SVM Classifier with MFCC and TEO Features for GDM-M and GDM-F and Each Interaction .......................................................... 256

Table 7-11 **Depression Classification Accuracy (%)**, Sensitivity and Specificity Using SVM with MFCC Extracted from Speech Signals with Various Levels of AWG Noise for GDM-F and GDM-M and Each Interaction .................................................................................. 259

Table 7-12 **Depression Classification Accuracy (%)**, Sensitivity and Specificity Using SVM with TEO Extracted from Speech Signals with Various Levels of AWG Noise for GDM-F and GDM-M and Each Interaction .................................................................................. 259

Table 7-13 **Depression Classification Accuracy (%)**, Sensitivity and Specificity Using SVM with Glottal Time-Domain (GTD) and Frequency-Domain (GFD) Features for GDM-M and GDM-F and Each Interaction .................................................................................. 260

Table 7-14 **Depression Classification Accuracy (%)**, Sensitivity and Specificity Using SVM with the Glottal Waveform Category Features (GW) (G-MFCC and G-TEO) for GDM-M and GDM-F and Each Interaction .................................................................................. 261

Table 7-15 **Depression Classification Accuracy (%)**, Sensitivity and Specificity Using SVM with Prosodic (P), Spectral (S), MFCC, TEO, G-MFCC and G-TEO Category Features in Addition to the Glottal Category Features (+G) for Male Subjects (GDM-M) for Each Interaction .......................................................... 263

Table 7-16 **Depression Classification Accuracy (%)**, Sensitivity and Specificity Using SVM with Prosodic (P), Spectral (S), MFCC, TEO, G-MFCC and G-TEO Category Features in Addition to the Glottal Category Features (+G) for Male Subjects (GDM-M) for Each Interaction) .................................................................................. 263

Table 7-17 **Difference in Adolescent Depression Classification Accuracy (%)** Between the Individual Acoustic Feature Set (S, P, MFCC, G-MFCC, TEO, G-TEO) and the Addition of the Glottal Category (Feature+G) .................................................................................. 264

Table 7-18 **Depression Classification Accuracy (%)**, Sensitivity and Specificity Using SVM with Prosodic (P), Glottal (G), MFCC, TEO, G-MFCC and G-TEO Category Features in Addition to the Spectral Category Features (+S) for Male Subjects (GDM-M) for Each Interaction) .................................................................................. 265
Table 7-19 Depression Classification Accuracy (%), Sensitivity and Specificity Using SVM with Prosodic (P), Glottal (G), MFCC, TEO, G-MFCC and G-TEO Category Features in Addition to the Spectral Category Features (+S) for Female Subjects (GDM-F) for each Interaction) ................................................................. 265

Table 7-20 Difference in Adolescent Depression Classification Accuracy (%) Between the Individual Acoustic Feature Set (G, P, MFCC, G-MFCC, TEO, G-TEO) and the Addition of the Spectral Category Feature (+S) ................................................................. 266

Table 7-21 Adolescent Depression Classification Accuracy, Sensitivity and Specificity Using SVM with Acoustic Feature Categories Prosodic (P), Spectral (S), Glottal (G), Cepstral (C), TEO (T) and Glottal Waveform (GW) Extracted from Their Parent’s Speech for Fathers (GDM-M) and Mothers (GDM-F) and Each Interaction ........ 268

Table 7-22 Ranking of Average Accuracy (%) Depression Classification Performance of Each Topic (EPI, FCI, PSI) and Gender (GDM-M and GDM-F) Using the ORI Database Speech for the Acoustic Feature Approach. The Shade of the Grey Cells Signify, from Lightest to Darkest: Sub-Category, Category, “+G” and “+S”. The Blue Shade Represents the Parent Features .......................................................................................................................... 272

Table 8-1 Ranking of the Depression Classification Performance of the Average Accuracy (%) of Each Topic (EPI, FCI, PSI) and Gender Model (GDM-M and GDM-F) Using the ORI-DB Speech Signals for the Acoustic Feature Approach and the LIFE Construct Annotations (positive, negative, neutral, silence) for the Influence Model Approach. The Dark Shaded Cells Denotes the Newly Proposed EIM Features, Lighter Shaded Cells Indicate the Speech Acoustic Features that Include New Features and Clear Cells are Speech Acoustics Used Previously .................................................................................................................. 287

Table 8-2 Summary of Current Literature That Most Closely Relate to the Acoustic Depression Detection Studies Presented in This Thesis. The Studies are Ranked from Best to Worst Based on the Top Feature Found in the Respective Studies. (* Denote Studies that Used ORI-DB) .......................................................................................................................... 290
LIST OF FIGURES AND ILLUSTRATIONS

Figure 1-1 Overview of the entire system diagram showing the conversation modeling system (CMS) used for depression detection requires emotion labels that could be manually or automatically generated from the emotion recognition (AER) or emotion prediction systems ..................................................8

Figure 2-1 General source-filter model for speech production, Y(z), where the excitation generator produces a source, X(z) either modeled by noise (unvoiced) or impulse train and glottal pulse model. The vocal tract model, V(z), and radiation model (lips), R(z), then filter the source.................................22

Figure 4-1 Flow chart outlining the main steps involved in each emotion recognition experimental setup including tasks related to pre-processing, feature extraction (glottal and speech waveform) and feature type (1D or 2D) for respective classifiers.................................................................68

Figure 4-2 General overview of emotion recognition training stage for N emotion models (binary or multiclass) including a training set, pre-processing, feature set extraction and classification system.................................................................69

Figure 4-3 General overview of emotion recognition test stage for N emotion models (binary or multiclass) including a test sample, pre-processing, feature set extraction and classification using the trained emotional models ..........69

Figure 4-4 Illustration of the voice activity detector from left to right, top to bottom (a-d) correspond to: (a) Entire speech signal (b) Concatenation of the extracted voiced frames (c) and (d) spectral-domain for original signal and after VAD........................................................................................................72

Figure 4-5 a) Mel-Scale versus frequency (Hz) scale conversion and b) triangular Mel filter-banks with non-linear separation on the frequency scale.................74

Figure 4-6 Wavelet Packet Tree (WPT) with three levels of decomposition using low pass filter denoted by L_i and the high pass filter by H_i. The filter coefficients for approximation and detail are the approximation and detail coefficients........................................................................................................77

Figure 4-7 Flow chart outlining the process of calculating the TEO_CB_AUTO_ENV coefficients from voiced speech frames. The derivation involves generating the TEO for the 17 sub-bands, found from the WPC, and then calculating the area under the autocorrelation contour.............78

Figure 4-8 Examples of 2D (34x34) image representations of acoustic feature sets as a contour (white) given for 12 MFCC coefficients (left) and 17 TEO coefficients (right). Where the x-axis represents the coefficient index and the y-axis represents the normalized range of coefficient values.................80

Figure 4-9 Illustration of linearly separable data to train a SVM with ‘+’ and ‘*’ indicating points for each class. The hyperplane, H, separates the two classes with a maximal margin, 2/||w||, between the parallel hyperplanes H_1 and H_2.
The data points that lie on the hyperplanes, denoted as circled points, are the support vectors.

Figure 4-10 Perceptron with \( n \) binary inputs, \( x_i \) and corresponding weights, \( w_i \), used to produce a binary output, \( z \), as a linear weighted sum of the input vector.

Figure 4-11 Artificial neuron structure with inputs, \( x_i \) and weights, \( w_i \), used in a non-linear weighted sum and an activation function, \( f(s) \), to produce an output, \( y \).

Figure 4-12 Neural Network architecture showing the connections of the artificial neuron nodes, which individually correspond to Figure 4-11, between the input, hidden and output layers.

Figure 4-13 DNN architecture representing \( L \) hidden layers (\( h \)), input and output layers are visible and based on the size of the features and the number of classes. The input neurons are weighted, \( W^{(L)} \), by vectors to generate visible, \( v \), outputs. The DNN structure is similar to the shallow NN however contains multiple hidden layers, \( h^{(L+1)} \).

Figure 4-14 Illustration of DBN layer interconnections of the weight matrix, \( W \), with visible layer units, \( v_m \), and the hidden units, \( h_n \) (left) and representation of the layer-wise training of a 3-layer DBN as an unsupervised RBM (right). Where the input to each RBM layer, \( v \), is the output of the previous with hidden layers denoted by \( h \).

Figure 4-15 Process of testing unknown sample, \( x_i \), with the general \( K \) channel OMC-DNN such that \( K \) independent feature sets are classified by individual SC-DNN with the results, \( \hat{y}_k(x_i) \), weighted by optimized weights, \( w_{k_{opt}} \), and the final class result, \( h_{xi} \), converted by the decision matrix.

Figure 4-16 Procedure for the general \( K \) channel OMC-DNN system to determine the optimized set of channel weights \( w_{k_{opt}} \) using the Simulated Annealing Genetic Algorithm Hybrid (SA/GA) optimization algorithm. The weights are iteratively updated, \( w_{k_{update}} \), until convergence based on the objective function, \( f_{obj}(w) \) that increases OMC-DNN accuracy.

Figure 4-17 Flow diagram outlining the process of Simulated Annealing (SA) including the criteria for acceptance of worse solutions and stopping criteria.

Figure 4-18 The Boltzmann’s acceptance probability function used in the classical Metropolis algorithm and simulated annealing. The temperature reduces with iteration number where \( T_i > T_2 \), the threshold, \( \zeta \), is a random number between 0 and 1.

Figure 4-19 Average accuracy of DNN using the four different 2D features (S-MFCC, G-MFCC, S-TEO, G-TEO). Results are supplied for each of the six binary systems for gender independent modeling (GIM).
Figure 4-20 Comparison of accuracy between gender independent (GIM) and gender dependent models (GDM) using DNN for the 6 different binary emotion recognition tasks. Results are supplied separately for the MFCC, TEO, G-MFCC and G-TEO.................................................................112

Figure 5-1 Average proportion (%) of time adolescents and parents spent in each construct state comparing conversations with depressed and non-depressed adolescents averaged for both genders and all interactions (EPI, FCI and PSI)132

Figure 5-2 Total average duration (seconds) consecutively spent in each construct state comparing conversations with depressed and non-depressed adolescents averaged over both genders and all interactions (EPI, FCI and PSI).................132

Figure 5-3 Generalized coupled HMM for $M$ chains for interaction between $t-1$ and $t$ with states, $S$ and hidden states, $x$ .................................................................136

Figure 5-4 Diagram of an $M$ chain general influence model shown in graphical form, illustrating the connection between and within chains, with continuous observations of state $S$ at times $t, t+1, ... , t+n$, ..........................................................................................139

Figure 5-5 Parent-Adolescent conversation represented by two coupled time chains as a general Influence Model (IM). Each node denotes the state a speaker is at time $t$. The system models the influence of each speaker on themselves (self-influence $\theta_{ii}$) and by others (cross-influence $\theta_{ij}$) with time delay $t-1$ between each transition.................................................................141

Figure 5-6 Illustration of Dynamic Emotional Influence Model (DEIM) depicting a dyadic parent-adolescent conversation as an interaction between Markov chains, with nodes denoted by the construct states (positive, negative, neutral silence) with arbitrary time delay ($t-n$) between transitions .........................143

Figure 5-7 Higher Order Emotional Influence Model (HOEIM) diagram of Markov chain interaction, depicting parent-adolescent conversation, with $N$ dependent time delays to model influence from multiple past construct states (positive, negative, neutral, silence) .................................................................148

Figure 5-8 A general framework for the conversation modeling system (CMS) approach to depression detection with Influence Model (IM) feature extraction (Influence Coefficients ($\Phi$) and Probabilities ($P$)) and machine learning training and testing/classification phases (SVM or NN) ...............148

Figure 5-9 Procedure to optimize number of hidden neurons in the NN such that the aim to find $H_{opt}$ that results in the highest average CV accuracy, $A_{avg}$ ..........151

Figure 5-10 Relative difference between the non-depressed and depressed adolescent’s autocorrelation function for the negative (left) and positive (right) types. Including three interactions: event planning interaction (EPI), family consensus interaction (FCI) and problem solving interaction (PSI). .....156

Figure 5-11 Relative difference between the non-depressed and depressed adolescent and parent sequence cross-correlation function for the negative
adolescents (left) and positive (right) types. Including three interactions: event planning interaction (EPI), family consensus interaction (FCI) and problem solving interaction (PSI).

Figure 5-12 Comparison of optimization methods, (Expectation Maximization (EM), Quasi-Newton (QN), Multivariate Linear Regression (MVR), Gradient Decent (GD), Simulated Annealing (SA) and Simulated Annealing/Genetic Algorithm Hybrid (SA/GA)), to learn the Emotional Influence Model (EIM) parameters using a delay of $t-1$. Performance is compared on the error rate (MSE) and CPU time (seconds) to learn the parameters, averaged over a portion each the ORI-DB.

Figure 5-13 Log-Likelihood ($LL$) comparison of different HOEIM implementations (MMM, ngram, KN-ngram) for each order $(N=1,...,5)$ and compared to the DEIM for delays of $t-n=1,...,10$.

Figure 5-14 Average MMM weights $\omega ij(nm)$ with time delay $nm$ between the current and the previous state; $\omega 11(nm)$–adolescent-adolescent, $\omega 22(nm)$–parent-parent, $\omega 12(nm)$–parent-adolescent, $\omega 21(nm)$–adolescent-parent.

Figure 5-15 Mixed Memory Markov Model (MMM) weights ($\omega 11(nm)$, $\omega 12(nm)$, $\omega 21(nm)$, $\omega 22(nm)$), with time delay $nm$ between the current and the previous state, generated for two different dyadic conversations (left and right).

Figure 5-16 Average DEIM self-influence ($\theta 11$ and $\theta 22$ (solid lines)) and cross-influence ($\theta 12$ and $\theta 21$ (dashed lines)) for time delays of $n=1,...,10$, comparing parameters of depressed (D (blue line)) and non-depressed (ND (red line)) adolescents (left) and their parents (right) during the EPI.

Figure 5-17 Average DEIM self-influence ($\theta 11$ and $\theta 22$ (solid lines)) and cross-influence ($\theta 12$ and $\theta 21$ (dashed lines)) for time delays of $n=1,...,10$, comparing parameters of depressed (D (blue line)) and non-depressed (ND (red line)) adolescents (left) and their parents (right) during the FCI.

Figure 5-18 Average DEIM self-influence ($\theta 11$ and $\theta 22$ (solid lines)) and cross-influence ($\theta 12$ and $\theta 21$ (dashed lines)) for time delays of $n=1,...,10$, comparing parameters of depressed (D (blue line)) and non-depressed (ND (red line)) adolescents (left) and their parents (right) during the PSI.

Figure 5-19 DEIM generated self Influence Coefficients (IC), $\theta ii$, for each time delay, $t-n$, comparing the self IC for each interaction (EPI, FCI, PSI) given for the non-depressed adolescents, $\theta 11_{ND}$, (left) and depressed adolescents, $\theta 11_{D}$, (right).

Figure 5-20 Average HOEIM self-influence ($\theta 11$ and $\theta 22$ (solid lines)) and cross-influence ($\theta 12$ and $\theta 21$ (dashed lines)) for orders $N=1,...,5$, comparing parameters of depressed (D (blue line)) and non-depressed (ND (red line)) adolescents (left) and their parents (right) during the EPI.
Figure 5-21 Average HOEIM self-influence ($\theta_{11}$ and $\theta_{22}$ (solid lines)) and cross-influence ($\theta_{12}$ and $\theta_{21}$ (dashed lines)) for orders $N=1,\ldots,5$, comparing parameters of depressed (D (blue line)) and non-depressed (ND (red line)) adolescents (left) and their parents (right) during the FCI.

Figure 5-22 Average HOEIM self-influence ($\theta_{11}$ and $\theta_{22}$ (solid lines)) and cross-influence ($\theta_{12}$ and $\theta_{21}$ (dashed lines)) for orders $N=1,\ldots,5$, comparing parameters of depressed (D (blue line)) and non-depressed (ND (red line)) adolescents (left) and their parents (right) during the PSI.

Figure 5-23 HOEIM generated self Influence Coefficients (IC), $\theta_{ii}$, for each order, $N$, comparing the self IC for each interaction (EPI, FCI, PSI) given for the non-depressed adolescents, $\theta_{11\_ND}$, (left) and depressed adolescents, $\theta_{11\_D}$, (right).

Figure 5-24 Average accuracy of NN classification with respect to the number of hidden nodes. Results averaged over delay lengths $n=1,2,\ldots,10$ separately for each of the interactions (EPI, FCI, PSI) and each node iteration (10,15,20,25,30,35 and 40).

Figure 5-25 Comparison of average depression classification accuracy using the DEIM features ($P_{11}, P_{12}, P_{21}, P_{22}, \theta_{11}, \theta_{12}, \theta_{21}, \theta_{22}$) at each delay ($t-n$) for the SVM, NN and Optimized NN.

Figure 5-26 ROC space of depression classification comparing three classification models SVM (red), 25 node Neural Network (blue) and optimized node NN (green) for EPI (left), FCI (middle) and PSI (right). Each point on the ROC space corresponds to the set time delay ($n=1,\ldots,10$) used to extract the DEIM feature set.

Figure 5-27 Framework of the feature selection method using a 2-stage mRMR filter to generate a ranking of top acoustic feature category used in a SVM wrapper stage to iteratively remove the lowest ranked feature to find the optimal accuracy.

Figure 5-28 ROC space of SVM depression classification comparing performance of features extracted from three HOEIM variants. The HOEIM-MMM, HOEIM-KNngram and 2-stage mRMR optimized KN-ngram feature sub-set are shown separately for EPI (left), FCI (middle) and PSI (right) where each point is an order ($N=1,\ldots,5$) of the respective EIM feature set.

Figure 5-29 Average depression classification accuracy, across the interactions (EPI, FCI, PSI), comparing the HOEIM feature extraction methods (HOEIM-MMM, HOEIM-KN-ngram and HOEIM-KN-ngram Optimized with mRMR) for each order.

Figure 5-30 Average SVM depression classification performance directly comparing the DEIM delay ($t-n$) with the HOEIM orders ($N$) for both the MMM and KN-ngram implementations and mRMR feature optimization.
Figure 6-1 2D random walk on a graph illustrating the direction the speaker (particle) can move into either the positive (up (+)), negative (down (-)), neutral (right (N)) or silent (left (S)) state. The example demonstrates a speaker that starts at the origin (0,0) denoted as position \( u \) and then moves up (0,1) to position \( v \) in a positive state (+) with a probability of \( p_{uv} \) and then moves right (1,1) to position \( w \) in a neutral state (N) with a probability of \( p_{vw} \).

Figure 6-2 2D Random Walk process from the origin (0,0) to the final position \((x_f,y_f)\) with phasor notation to represent the magnitude of distance travelled, \(|D|\), and the drift being defined by the angle, \( \theta \), travelled.

Figure 6-3 General process of a Non-linear Autoregressive model (NAR) such that the output, \( y(t) \), is a function of it’s own past states, \( y(t-1),...y(t-n) \).

Figure 6-4 General process of a Non-linear Autoregressive model with external input (NARX) describing the output, \( y(t) \), as a function of its own past states, \( y(t-1),...y(t-n) \), and the past states of the external input, \( u(t-1),...u(t-m) \).

Figure 6-5 NAR Neural Network architecture, adapted from [255], illustrating estimated output, \( \hat{y}(t) \), is based on the time-delay line (TDL), the corresponding weights, \( w \), and bias, \( b \), parameters of the NN along with hidden layer and output layer activation functions, \( f \).

Figure 6-6 NARX-NN architecture, adapted from [255], illustrating estimated output, \( \hat{y}(t) \), based on time-delay line (TDL) of the output and external input, \( u(t) \), with corresponding weights, \( w \), and bias, \( b \), parameters and activation functions, \( f \).

Figure 6-7 Activation functions, \( f \), used for the hidden and output layers illustrating a tan-sigmoid (left) and a linear activation function (right).

Figure 6-8 Feed forward NARX architectures used to estimated a sequence, \( \hat{y}(t) \), with a tapped time delay line (TDL) of past states of an external input, \( u(t) \), and past states of the estimated output, \( \hat{y}(t) \), fed back in a closed loop (left) or using actual output, \( y(t) \), in an open loop (right) replicated from [255].

Figure 6-9 NAR Network implemented with ANFIS using past states of \( y \) at the input to the \( K \) membership functions \( (A_k) \) to generate the estimated output, \( \hat{y} \), as the normalized weighted, \( wk \), summation of each of the membership function outputs, \( f_k \).

Figure 6-10 NARX implemented with ANFIS using past data of \( y \) and external sequence \( u \) at the \( K \) input membership functions \((A_k \text{ and } B_k)\) to estimate the next state, \( \hat{y} \), as the normalized weighted, \( wk \), summation of each output MF result, \( f_k \).

Figure 6-11 Average relative error (%) of the Higher Order Random Walk (HORW) (filled bar) in comparison to the Interacting Random Walk (IRW) (hashed bar) for orders 1-5 and each iteration length (10, 25, 50 and 250 steps).
Figure 6-12 Average relative difference of distance $|D_{RMS}|$ and angle $\theta$ between real data and the random walks of each order for 25 (left) and 250 steps (right) .................................................................210

Figure 6-13 Representation of a random walk on a 2D graph of a speaker with positive, negative, neutral, silence in each direction with defined quadrants that characterize emotional drift. Each quadrant corresponds to information regarding the leading sign (positive or negative) of the affect and the dominance based on silence .......................................................211

Figure 6-14 Three random walks for 250 steps comparing to the actual data in blue and another three examples (RW path 1, path 2 and path 3) showing different emotional drift in the x and y direction .......................................................212

Figure 6-15 250-step random walks showing three different emotional drifts, negative-silent (green), positive-silent (black), positive-neutral (red).................212

Figure 6-16 2D graph representation of the RWs comparing the emotional drift of the 29 depressed (left) and 34 non-depressed (right) adolescents ............213

Figure 6-17 Summary of the speaker and class independent (SI-CI) setup giving the average accuracy (%), across three topics (EPI, FCI and PSI), comparing each order of the NAR-NN and NARX-NN affect prediction models ..........215

Figure 6-18 Average affect prediction accuracy (%), across topics (EPI, FCI, PSI), of speaker independent model (SI) comparing independence (CI) and class dependence (CD) with depressed (CD-D) and non-depressed (CD-N) adolescent models for both NARX-NN (left) and NAR-NN (right) models.....216

Figure 6-19 Average affect prediction accuracy (%), across topics (EPI, FCI, PSI), of class independent model (CI) comparing speaker independence (SD) and speaker dependence (SD) models for both NARX-NN NAR-NN models....217

Figure 6-20 Average affect prediction accuracy (%) for NAR-NN (left) and NARX-NN (right) using class dependent modeling comparing speaker independence and dependence (SI-CD and SD-CD) for Depressed (D) and Non-depressed (N).................................................................218

Figure 6-21 Average accuracy (%), across order and interactions, for the Speaker Dependent Class Dependent (SD-CD-D and SD-CD-N) NAR and NARX implemented with NN and ANFIS .................................................................220

Figure 7-1 Overall system diagram for acoustic depression recognition training and testing phases including main steps of preprocessing, feature extraction and SVM modeling ........................................................................223

Figure 7-2 Spectral Subtraction process used to estimate the clean speech signal, $x'(m)$, from the noisy speech, $y(m)$, using an estimation of the noise, $n(m)$ ......224

Figure 7-3 Diagram of cross-talk from two speakers (sources), $s_n(t)$, that are involved in the mixing process defined by a weight, $h_{nm}(t)$, creating mixed signals $s_{nm}(t)$, that are summed at the microphone, $x_n(t)$..................................................225
Figure 7-4 Diagrams of (a) top- Time-domain speech frame (b) middle- Glottal waveform estimate using TKK Aparat toolbox (c) bottom- Glottal flow derivative. Including the relevant phases and parameters required to generate the glottal time and frequency parameters .................................................. 237

Figure 7-5 Illustration showing the relative difference of the average spectral roll-offs between the non-depressed (ND) and depressed (D) subjects for a range of cut-off points \((k=5\%-95\%)\) included for male and female adolescents in EPI, FCI and PSI ................................................................. 248

Figure 7-6 Framework of the feature selection method using a 2-stage mRMR filter to generate a ranking of top acoustic feature category used in a SVM wrapper stage to iteratively remove the lowest ranked feature to find the optimal accuracy ................................................................. 250

Figure 7-7 Accuracy (%) of SVM GDM-M depression classification with 1 to 19 roll-off feature coefficients kept based on ranking of MIQ from the mRMR filter .................................................................................................. 251

Figure 7-8 Accuracy (%) of SVM GDM-F depression classification with 1 to 19 roll-off feature coefficients kept based on ranking of MIQ from the mRMR filter .................................................................................................. 251

Figure 7-9 Average accuracy (%) of the three interactions comparing the performance of the AWGN SNR levels for the MFCC and TEO features and each GDM .................................................................................................. 258

Figure 7-10 Average depression classification accuracy (%), across the three interactions (EPI, FCI, PSI), comparing the performance of MFCC and TEO derived from speech and glottal waveform in both GDM-M and GDM-F ...... 262

Figure 7-11 Average accuracy (%) of adolescent depression classification, for all interactions, comparing the “+G” parameters for the GDM-M and GDM-F .... 263

Figure 7-12 Average accuracy (%) of adolescent depression classification, for all interactions, comparing the “+S” parameters for the GDM-M and GDM-F .... 266

Figure 7-13 Average accuracy (%) of adolescent depression classification, for all interactions, comparing the feature categories (S, G, P, T, C, GW) from both the adolescents and parents speech for GDM-M and GDM-F ....................... 269
LIST OF ABBREVIATIONS

The following table provides a limited list of the most relevant and frequent abbreviations and acronyms used within this thesis with the corresponding definition and page of first appearance.

<table>
<thead>
<tr>
<th>Abbreviations</th>
<th>Definition</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>AER</td>
<td>Automatic Emotion Recognition</td>
<td>20</td>
</tr>
<tr>
<td>ANFIS</td>
<td>Adaptive Neuro fuzzy Interference System</td>
<td>14</td>
</tr>
<tr>
<td>AWGN</td>
<td>Additive White Gaussian Noise</td>
<td>243</td>
</tr>
<tr>
<td>C</td>
<td>Cepstral Category</td>
<td>228</td>
</tr>
<tr>
<td>CMS</td>
<td>Conversation Modeling System</td>
<td>iv</td>
</tr>
<tr>
<td>D</td>
<td>Depressed</td>
<td>154</td>
</tr>
<tr>
<td>DAP</td>
<td>Discrete all-pole model</td>
<td>73</td>
</tr>
<tr>
<td>DBN</td>
<td>Deep Belief Network</td>
<td>32</td>
</tr>
<tr>
<td>DEIM</td>
<td>Dynamic Emotional Influence Model</td>
<td>iv</td>
</tr>
<tr>
<td>DNN</td>
<td>Deep Neural Network</td>
<td>13</td>
</tr>
<tr>
<td>DSM-IV</td>
<td>Diagnostic and Statistical Manual of Mental Disorders</td>
<td>62</td>
</tr>
<tr>
<td>EIM</td>
<td>Emotional Influence Model</td>
<td>iv</td>
</tr>
<tr>
<td>EM</td>
<td>Expectation Maximization Algorithm</td>
<td>82</td>
</tr>
<tr>
<td>EMO-DB</td>
<td>Berlin Emotional Speech Database</td>
<td>12</td>
</tr>
<tr>
<td>EPI/FCI/PSI</td>
<td>Event Planning/Family Consensus/ Problem Solving Interactions</td>
<td>63</td>
</tr>
<tr>
<td>F0</td>
<td>Fundamental Frequency</td>
<td>24</td>
</tr>
<tr>
<td>G</td>
<td>Glottal Category</td>
<td>228</td>
</tr>
<tr>
<td>G-MFCC</td>
<td>Glottal Waveform MFCC</td>
<td>iii</td>
</tr>
<tr>
<td>G-TEO</td>
<td>Glottal Waveform TEO</td>
<td>iii</td>
</tr>
<tr>
<td>GDM (GDM-F/GDM-M)</td>
<td>Gender Dependent Models (Female/ Male)</td>
<td>107 (111)</td>
</tr>
<tr>
<td>GFD</td>
<td>Glottal Frequency Domain</td>
<td>238</td>
</tr>
<tr>
<td>GiM</td>
<td>Gender Dependent Model</td>
<td>107</td>
</tr>
<tr>
<td>GMM</td>
<td>Gaussian Mixture Model</td>
<td>17</td>
</tr>
<tr>
<td>GTD</td>
<td>Glottal Time Domain</td>
<td>238</td>
</tr>
<tr>
<td>GW</td>
<td>Glottal Waveform Category</td>
<td>228</td>
</tr>
<tr>
<td>HOEIM</td>
<td>Higher Order Emotional Influence Model</td>
<td>iv</td>
</tr>
<tr>
<td>HOEIM-KNngram</td>
<td>Modified Kneser-Ney Smoothed Ngram implemented HOEIM</td>
<td>iv</td>
</tr>
<tr>
<td>HOEIM-MMM</td>
<td>Mixed Memory Markov Model implemented HOEIM</td>
<td>iv</td>
</tr>
<tr>
<td>HOEIM-Ngram</td>
<td>Ngram implemented HOEIM</td>
<td>145</td>
</tr>
<tr>
<td>HORW</td>
<td>Higher Order Random Walk</td>
<td>14</td>
</tr>
<tr>
<td>IAIF</td>
<td>Iterative adaptive inverse filtering algorithm</td>
<td>73</td>
</tr>
<tr>
<td>IC</td>
<td>Influence Coefficients (i.e. $\theta_a$ and $\theta_d$)</td>
<td>135</td>
</tr>
<tr>
<td>IM</td>
<td>Influence Model</td>
<td>iv</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>IRW</td>
<td>Interacting Random Walk</td>
<td>14</td>
</tr>
<tr>
<td>LIFE</td>
<td>Living In Family Environment Coding System (Annotations)</td>
<td>iii</td>
</tr>
<tr>
<td>LL</td>
<td>Log-Likelihood</td>
<td>82</td>
</tr>
<tr>
<td>LLD</td>
<td>Low-level Descriptors</td>
<td>24</td>
</tr>
<tr>
<td>LOGE</td>
<td>Log Energy</td>
<td>29</td>
</tr>
<tr>
<td>MDD</td>
<td>Major Depressive Disorder</td>
<td>37</td>
</tr>
<tr>
<td>MFCC</td>
<td>Mel-Frequency Cepstral Coefficients</td>
<td>26</td>
</tr>
<tr>
<td>MMM</td>
<td>Mixed Memory Markov Model</td>
<td>143</td>
</tr>
<tr>
<td>MSE</td>
<td>Mean Square Error</td>
<td>90</td>
</tr>
<tr>
<td>NAR</td>
<td>Non-Linear Autoregressive Model</td>
<td>14</td>
</tr>
<tr>
<td>NAR-AFIS</td>
<td>NAR implemented with ANFIS</td>
<td>204</td>
</tr>
<tr>
<td>NAR-NN</td>
<td>NAR implemented with Neural Network</td>
<td>56</td>
</tr>
<tr>
<td>NARX</td>
<td>Non-Linear Autoregressive eXogenous Model</td>
<td>14</td>
</tr>
<tr>
<td>NARX-AFIS</td>
<td>NARX implemented with ANFIS</td>
<td>204</td>
</tr>
<tr>
<td>NARX-NN</td>
<td>NARX implemented with NN</td>
<td>56</td>
</tr>
<tr>
<td>ND</td>
<td>Non-depressed/Control</td>
<td>154</td>
</tr>
<tr>
<td>NN</td>
<td>Neural Network</td>
<td>17</td>
</tr>
<tr>
<td>OMC-DNN</td>
<td>Optimized Multichannel DNN</td>
<td>13</td>
</tr>
<tr>
<td>ORI-DB</td>
<td>Oregon Research institute Depression Audio-Visual Database</td>
<td>iii</td>
</tr>
<tr>
<td>P</td>
<td>Prosodic Category</td>
<td>228</td>
</tr>
<tr>
<td>$P_{ii} {P_{11}, P_{22}}$</td>
<td>Self Transition Probability Matrix (Adolescent, Parent)</td>
<td>140</td>
</tr>
<tr>
<td>$P_{ij} {P_{12}, P_{21}}$</td>
<td>Cross-transition Probability Matrix (Adolescent, Parent)</td>
<td>140</td>
</tr>
<tr>
<td>PSD</td>
<td>Power Spectral Density</td>
<td>41</td>
</tr>
<tr>
<td>RBM</td>
<td>Restricted Boltzmann Machine</td>
<td>32</td>
</tr>
<tr>
<td>RW</td>
<td>Random Walk</td>
<td>14</td>
</tr>
<tr>
<td>S</td>
<td>Spectral Category</td>
<td>228</td>
</tr>
<tr>
<td>SA/GA</td>
<td>Hybrid Simulated Annealing/Genetic Algorithm</td>
<td>96</td>
</tr>
<tr>
<td>SC-DNN</td>
<td>Single Channel DNN</td>
<td>96</td>
</tr>
<tr>
<td>SD-CD-(D/ND)</td>
<td>Speaker Dependent - Class Dependent Model</td>
<td>208</td>
</tr>
<tr>
<td>SD-CI</td>
<td>Speaker Dependent - Class Independent Model</td>
<td>208</td>
</tr>
<tr>
<td>SI-CD-(D/ND)</td>
<td>Speaker Independent - Class Dependent Model</td>
<td>208</td>
</tr>
<tr>
<td>SI-CI</td>
<td>Speaker Independent - Class Independent Model</td>
<td>208</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machine</td>
<td>17</td>
</tr>
<tr>
<td>T</td>
<td>Teager-based Category</td>
<td>228</td>
</tr>
<tr>
<td>TEO</td>
<td>Teager Energy Operator</td>
<td>26</td>
</tr>
<tr>
<td>TEO-CB-AUTO-ENV</td>
<td>Area under normalized TEO autocorrelation envelope in critical band</td>
<td>78</td>
</tr>
<tr>
<td>VAD</td>
<td>Voice Activity Detector</td>
<td>70</td>
</tr>
<tr>
<td>$\theta_{ii} {\theta_{11}, \theta_{22}}$</td>
<td>Self Influence Coefficients (Adolescent, Parent)</td>
<td>140</td>
</tr>
<tr>
<td>$\theta_{ij} {\theta_{12}, \theta_{21}}$</td>
<td>Cross Influence Coefficients (Adolescent, Parent)</td>
<td>140</td>
</tr>
</tbody>
</table>
Chapter One:

INTRODUCTION

1.1 Preview

The objectives of this introductory chapter are to provide a detailed background statement, in Section 1.2, discussing information related to depression specifically in relation to parent-adolescent conversations and the issues to current diagnosis techniques. An outline of problems related to adolescent depression and detection provides reasoning for the necessity for a new perspective on the work. Moreover, this chapter delivers the aims, scope, limitations and contributions in Sections 1.3-1.5 and the structure of the remaining chapters of this dissertation in Section 1.6.

1.2 Background and Problem Statement/Description

Clinical depression is an affective disorder consisting of emotional disturbances, prolonged phases of excessive sadness and reduced emotional expression [449]. Those with a major depressive disorder often experience decreased motivation, lack of energy, fatigue, loss of appetite, low self-esteem and irritability [6]. Depression is a debilitating disorder that greatly impairs a person's ability to function, which can jeopardize social, emotional, educational and working capabilities.

Mental illness has a large economic impact with depression associated expenses costing Australia $14.9 billion annually, including $600 million directly for treatments, and over 6 million working days lost annually [128].
Depressive disorders significantly contribute to disability as the leading cause of non-fatal disability in Australia (24%) [381]. Depression is the third highest disease burden in Australia accounting for 13.3% [381], the third leading cause of global burden of disease [127], by 2020 will be second [8] and highest by 2030 [126][127].

In 2001 the World Health Organization (WHO) estimated 121 million people worldwide were affected by depression [8][506] and just over a decade later estimated at 350 million [127]. The Australian Bureau of Statistics (ABS) states over 40% of people aged 16 to 85 experience a form of mental illness during their lifetime [378]. Specifically, major depression is the most prevalent disorder with a lifetime risk of 5-12% for males and 10-20% for females [476]. In any one year 20% will experience a mental illness [377] and 6.3% will get treatment for clinical depression [129].

In Australia mental illness is most prevalent in 16-24 year olds with 26% experiencing a disorder every year [379] including anxiety disorders (14%), depressive disorders (6%) and substance abuse disorders (5%) [380][1]. Since the 1970s there has been a substantial increase in the prevalence of adolescents with depression [448][132]. Within a 12-month period 14-30% of young females and 13-17% of males are affected by depression [131]. It is suggested by the age of 18 one in five are likely to experience a diagnosable depressive episode [4][5][131][133].

Depression commonly first appears in adolescence or early adulthood, emerging for half of all lifetime cases by the age of 14 and 3/4 by the age of 24 [2][3]. During adolescence depression is associated with energy loss, social withdrawal, disruptive behaviors and often leads to life-long reoccurrence into adulthood [335].

At worst depression can lead to suicide, a serious public health issue with lasting harmful effects on sufferers, family, friends and community [483]. The ABS has found suicide accounts for a small proportion (1.6%) of deaths in Australia [383].
More than 180 suicides each month relate to depression [472], with at least six Australians committing suicide each day and 30 more attempts [384]. Depression is a leading cause of suicide accounting for around two-thirds of suicides each year [413].

It is estimated that the suicide risk in those with depression is 30 times higher compared to a healthy person [414]. Men are a high risk of suicide, accounting for over three-quarters of suicide deaths [384]. Indigenous Australians and those in rural or remote areas and children are at a greater risk [384].

The increased prevalence of depression during adolescence has been associated with a rise in suicide and suicide attempts [132][131][133]. In Australia, suicide is the leading cause of death for young people aged 15-24 [383], with the fourth highest youth suicide rate in the industrial world [133].

The lifetime risk of suicide with untreated depression is 20% [385] and is reduced to below 1% with treatment [386]. Once depression is detected 70% to 80% of people are successfully treated with positive results and improve their lives [387].

Although treatment can be effective only 20% with a depressive illness seek treatment [387]. The WHO reports the main diagnosis barriers are a lack of resources and health care providers. The availability and access of health care is a problem, approximately 65% with mental illness do not have access treatment [381][382]. The WHO found 76%-85% of people with severe mental disorders receive no treatment in low and middle-income countries and 35%-50% in high-income countries [130].

Certain demographics have a higher risk of not seeking help such as males with an estimated 72% not treating mental disorders [384] and up to 80% of Americans relying on the Internet as an alternative to professional health diagnosis [370]. There is a perceived stigma with mental illness, evident by the fact those accessing treatment for mental illness is only half of those with physical issues [382].
Depression in young people is especially difficult to recognize, and symptoms can be easily untreated, undiagnosed or unrecognized during initial appearance [9]. Difficulties in adolescent depression detection is due to patients not providing reliable or accurate information about their own level of symptomatology and may not express feelings or not aware of their falling themselves. Moreover, there is a general perception that it is uncommon to suffer from mental illness at a young age. Detection during adolescence is the main hurdle to overcome if under-treated it is likely to lead to life-long reoccurrence [335] and difficulties in adulthood diagnosis [131].

Currently, from a conventional psychological strategy depression diagnosis almost entirely relies on professional observations and evaluations based on clinical interviews/questionnaires from the patient, family and caregivers. Professional evaluation is dependent on clinical judgment and diagnostic methods (DSM-IV [22], Hamilton Rating Scale for Depression [444], Quick Inventory of Depressive Symptomatology [495]) [1]. These techniques are subjective qualitative assessments that depend on the training, skill set, experience and judgment of the practitioner. A common practice is self-reports (i.e. Beck Depression Index [452]), hindered by the reliance on patient’s depressive symptoms experience that are idiosyncratic [437].

The increased prevalence of depression [127] associated economic costs [126], disease/disability burden [128] and negative impact on sufferers [449][6]; particularly increased suicide rates [413] have become an issue. Depression and suicide rates are arguably worse in youth [379][383][133], and without intervention is a life-long problem [2][3][335]. Research has suggested detection and treatment are beneficial to potential recovery and relapse likelihood [386][387] especially with early detection during adolescence [131].
The problems are with current subjective diagnosis techniques and sufferers not seeking treatment [130][387][381][382] and detection difficulties are more prominent in adolescents [9][131]. Therefore, a main challenge is to reliably detect clinical depression, especially early on in adolescence.

The motivation for this research is to improve upon current diagnostic techniques with an accessible, efficient, affordable, natural and non-invasive objective automatic depression detection system. Ultimately, the goal is to create an automatic software application that would improve detection rates of depression and remove subjective bias. Requiring only a microphone and recording device, computer or telecommunications (mobiles), improves access and discreetness to increase the number of people willing and able to seek treatment.

The tool would provide initial mass screening but should not and would not replace mental health practitioners but would provide initial indicators and be followed by a standard full clinical evaluation. This could provide assistance with additional qualitative and quantitative measures to support and complement their assessment. A potential is to automatically analyze patient’s speech during a psychologist interview to objectively monitor depression and treatment.

The ultimate goal would be to eventually extend the system towards general mental health detection (stress, schizophrenia, depression) to detect disorders, monitor symptoms, severity and suicide risk. This could be applied to prevention center hotlines (i.e. Lifeline, Suicide Call Back Service) and telemedicine for remote assessment. Those with stressful and life risking jobs such military, pilots or doctors could be analyzed from conversations (radio transmissions). Another eventual purpose could be applied in forensic psychological assessments in legal issues [459].
Clinical depression is clearly associated with emotion and is considered an affect (emotion) regulation disorder [24], associated with emotional disturbances, excessive sadness and anger [445][449]. Depression has a strong correlation with increased expressed emotions, mood changes, and stress [21][100][16][310][311].

Emotions play an important role in adolescent depression and shows indications during parent-child interactions [24][135][136]. Depressed adolescents demonstrate a resistance to emotional change [87][88] and difficulties in regulation of negative emotions [24][137].

Depressed adolescents struggle to identify their parent’s mood and often misinterpret emotions [320]. Depressed adolescents over-perceive aggressive behavior and under-perceive positive affect [320] and have the tendency to misattribute happy, neutral and sad expressions as anger [320][319].

Existing psychological studies have found evidence that the quality of family interactions is correlated to adolescent depression and a critical factor in depression symptoms [24][134][135][136][137][145][157][320][144][319]. Depressed children are more likely to experience an adverse family environment [145][157], characterized by elevated levels of negative interactions and an absence positive [136].

Negative reciprocal parent-children process reinforces the children’s depressive behaviors [135][137][24][144]. Even after depression remits, chronic maladaptive family-based patterns of emotional interactions persist, and potentiate recurrence of depression [146]. It is important to detect depression and prevent the relapse of depression by educating families to change maladaptive patterns and reduce negative interactions that reinforce depression [336][135][137][144][145].
Considering the link between emotion and depression it would be expected that analysis of emotional characteristics could be applied to detect clinical depression. Additionally, the association of adolescent depression and family interaction in terms of emotions could have discriminating factors for depressed and non-depressed adolescents.

The scope of this research develops a new adolescent depression detection system by proposing a conversation modeling system (CMS), established to provide vital information from emotional interaction patterns with parents.

The CMS determines qualitative risk factors evident in depressed adolescents that can provide a basis for behavioral therapies to reverse maladaptive behaviors to treat depression. This could include helpful indicators to assist in Interpersonal Psychotherapy (IPT), an empirically validated treatment for depression [447]. Furthermore, the new computational methodology can provide a means of depression detection from quantitative parameters with a classification system.

The CMS depression detection approach can be compared with classical acoustic features extracted from the adolescent combined with machine learning. Additional investigations could determine if adolescent depression can be detected from their parent’s speech. It could provide more discrete diagnosis, and it could be beneficial if children do not want to cooperate with diagnostic procedures.

The purpose of this research was to investigate characteristics of depression, based on emotions, conversations and speech parameters and use it to develop an objective system that assists in diagnosis and monitoring of depression.

This introduction chapter contains details on the problem of detecting clinical depression in adolescents, the aims and scope of the project, a description of contributions of this research and an overview of the rest of this dissertation.
1.3 Thesis Aims and Research Questions

This research has been motivated by the need for an automatic system for depression detection and analysis. The aim is to develop objective measures of risk factors in depression and diagnosis on a mass-scale from a classical acoustic speech feature approach or from new conversation models. Depression is related to emotion and regulation of emotions [21][100][16][310][311][445] and is correlated to emotional conversation patterns in parent-adolescent interactions [24][135][136].

Figure 1-1 Overview of the entire system diagram showing the conversation modeling system (CMS) used for depression detection requires emotion labels that could be manually or automatically generated from the emotion recognition (AER) or emotion prediction systems. Figure 1-1 gives the system diagram outlining how the individual research components could be integrated together including the automatic emotion recognition (AER), conversation modeling system (CMS) and emotion prediction. The general concept is to automatically analyze emotional annotations of conversations in relation to depression detection using in the CMS. Learned CMS parameters characterizing conversation could be fed to the next component to improve emotion prediction.

The required emotion labels could be automatically generated from AER using speech or from the emotion prediction system using past emotional states. AER performance degrades in speakers with clinical depression [235][143] and could be improved using a front end to detect depression and then depression dependent AER models. Automatic systems are erroneous and these components are only investigated as independent systems with manual annotations used in the CMS instead.
The research aims and questions are separated into depression (speech and conversation modeling) and emotion (recognition and prediction) categories. The thesis aimed to provide answers to the following research questions:

**Emotion recognition and prediction aims and research questions:**

Aim 1) Research and develop a new emotion recognition system from speech signals:

Q1. What features are most effective for binary and multiclass AER?

Q2. How does a new deep learning approach compare to benchmark classifiers?

Q3. Can an optimized multichannel classifier improve emotion recognition performance compared to the benchmark systems?

Aim 2) Research and develop methods for predicting/forecasting sequences of emotional states in dyadic conversations:

Q1. Is it possible to predict/forecast emotion of adolescents in conversations based on a memory of past states?

Q2. Do higher order memories improve the performance of affect prediction?

Q3. How to incorporate dependencies/influences between speakers towards improving affect prediction.
**Depression detection aims research questions:**

Aim 3) Research and develop a new probabilistic, generative or statistical method for a conversation modeling system (CMS):

Q1. What is the most capable and efficient model of emotional interactions within conversations?

Q2. How to optimally fit the new conversation modeling system, Influence Model, into the data efficiently?

Aim 4) Apply the new conversation modeling system to parent-child conversations in families with depressed and non-depressed children:

Q1. Can the CMS provide psychologically valid qualitative interoperations and quantitative statistical significance?

Q2. What CMS parameters provide the best depression discriminators?

Q3. How to most efficiently automatically detect depression from the CMS?

Aim 5) Research and apply new methodologies for speech depression detection:

Q1. Is it possible to detect adolescent depression from speech parameters during family conversations?

Q2. What acoustic features show significant differences and provide discriminating risk factors related to depression in adolescents?

Q3. What is the optimal combination of features and classification setup towards achieving the most efficient depression detection performance?

Q4. Can depression in adolescents be detected by analyzing the acoustic speech features from their parent’s speech during family interactions?

Q5. How does acoustic depression detection compare to the newly proposed conversation modeling system approach?
1.4 Thesis Scope and Limitations

One of the main areas of investigation in this dissertation was to research links between adolescent depression and acoustic speech properties of adolescents and their parents. Another important area was the analysis and prediction of emotional patterns between parents and children in depressed and non-depressed family environments using forecasting models.

The main challenges of this research were related to difficulties in acquiring suitable large database of speech samples representing depressed and control subjects with corresponding annotations. For conversation analysis, acoustic modeling and prediction, a database collected by the Oregon Research Institute (ORI) was used and was not specifically collected for this research. The limitations to the scope of this thesis were imposed as follows:

1) A restricted age range of adolescents (14-18 years) was obtained during the collection with the age demographic when depression symptoms are common.

2) There are many forms of mental illnesses and different types of depression disorders. However, this research is restricted to Major Depression, which is the most common depressive disorder in adolescence.

3) The tests are limited to English speakers selected from a specific “high risk for depression” and low socioeconomic background population of the Oregon state of the U.S.A. This increases the potential for language dependency of the proposed speech analysis methods.
4) The conversation analysis research is restricted to dyadic conversations parent-adolescent conversations and omits conversations between adolescents and both their parents.

5) The conversational data contained 29 depressed (5 male 24 female) and 34 controls individual (24 female 10 male). This shows that, the genders of the dyadic conversations were heavily biased towards the mother-daughter teams with very few fathers participating. These numbers reflected the demographic distribution of people who agreed participate in the data collection process.

6) The original emotional annotation labels of the ORI data include 10 different emotional states. To reduce the complexity of the modeling and interoperation the number of labels was limited to only four construct states (positive emotion, negative emotion, neutral and silence). Further explanations of the annotation and reasoning for constructs are provided in Sections 3.3.4 and 5.2.

7) Practical applications of the proposed methods into designing new therapies or assessments of outcomes of existing therapies fall outside the project scope.

8) Acoustic emotion recognition tests described in this thesis were conducted using only one database – the Berlin Emotional Speech Database (EMO-DB). This limitation reduces the value of the validation process and further tests should be conducted in the future with other databases.

9) Furthermore, the CMS is restricted to manually annotated sequences of emotions (LIFE) and a completely automatic system with integration of an automatic emotion recognition (AER) system is beyond the scope of this work
1.5 Thesis Contributions

This thesis presents a series of systems that provide important contributions to the research towards emotion recognition (acoustic), depression detection (conversation modeling), adolescent depression detection (acoustic) and emotional interaction prediction (modeling). The thesis addresses an important computational open question about the relation of the two components of speech with the thesis contributions summarized as follows:

• **Automatic Emotion Recognition Contributions**

1) It was shown Deep Neural Networks (DNN) could more accurately classify emotions, in comparison to classical shallow NN, SVM and GMM classifiers, from new 2D image representations of acoustic feature vectors. The proposed representation of features (MFCC, TEO, G-MFCC, G-TEO) as contours in images is the first of its kind.

2) A comparison based on the EMO-DB has shown that the DNN combined with the new 2D features was able to significantly improve the AER performance compared to benchmark classifiers using either 2D or standard feature vectors.

3) It was observed that extracting TEO based parameters from the glottal waveform rather than directly from the speech signal improves the performance of the DNN, as well as other AER approaches.

4) This research proposed a new weighted optimized multichannel Deep Neural Network (OMC-DNN). It was found that the new OMC-DNN outperformed the single channel DNN for binary and multiclass setups.
• Prediction of Emotional Interactions Contributions

1) A new Higher Order Random Walk (HORW) approach to the prediction of speaker’s emotional states was introduced. It was shown that, the proposed higher order approach is more efficient in conversational affect prediction compared to the traditional first order models.

2) A new interacting random walk (IRW) method was introduced. The proposed approach included the weighted influence of an external speaker (conversational partner). This influence was estimated using the HOEIM influence coefficients (IC) proposed in this thesis. The new IRW generally improved the accuracy of affect prediction compared to the simple biased random walk (RW).

3) New non-linear autoregressive models (NAR) were proposed to predict adolescent affect and improved further using an external input (NARX) in parent-adolescent conversations. It was shown that the higher order models (i.e. including longer memory of the previous states), improved prediction accuracy and the NARX improved compared to respective NAR model.

4) It was shown that the NAR(X) approach implemented with the adaptive neuro fuzzy interference system (ANFIS) outperformed the NAR(X) implementations based on the basic NN.
• **Conversation Modeling System (CMS) Depression Detection Contributions**

A new approach to depression detection was introduced based on an emotion conversation modeling system (CSM) as an extension to the existing Influence Model (IM). This is used to characterize emotional intra- and inter-speaker influence coefficients in conversations and discriminate between depressed or non-depressed subjects. This thesis makes contributions to the analysis and understanding of emotional conversations in the context of depression as follows:

1) New Emotional Influence Model (EIM) facilitated more in depth interoperation in qualitative analysis in risk factors in depression. Generated by the replacement of traditional 2-state (speech and silence) analysis to 4 states including emotions.

2) The EIM with a constant IC value was extended into a new dynamic Influence Model (DEIM) with ICs given as a function of time.

3) The first order EIM was extended into Higher Order Emotional Influence Model (HOEIM) showing an improved fit of data compared to first order approach.

4) The proposed CMS provided new insights into conversation dynamics with quantitative and qualitative characteristics particularly suitable for potential depression analysis, detection and therapies.

5) It was demonstrated that DEIM and HOEIM parameters were statistically significant discriminators between depressed and control adolescents with higher order models outperforming the first order approaches.

6) The qualitative measures derived from descriptions of parameters from the proposed models have shown consistency with psychological interoperations.

7) It was shown that the new 2-stage mRMR optimized feature sub set of HOEIM parameters and an optimized NN for the DEIM parameters can provide particularly high accuracy of depression detection.
• Acoustic Speech Depression Detection Contributions

Traditional methods of automatic depression detection that rely on speech acoustics depend on the recording quality as well as, gender, age cultural background and language of the speaker.

Conventional methods of depression diagnosis based on acoustic speech parameters were compared with diagnosis based on parameters describing the newly proposed conversation models (DEIM and HOEIM). Unlike the classical approaches, the model-based classification was more likely to take into account some of the influences produced by gender, age or cultural backgrounds of speakers. The thesis contributions in this area include:

1) In particular an investigation of the usefulness of the glottal waveform in depression detection was conducted. It found that depression detection was generally improved when features parameters were extracted from the glottal waveform rather than the speech signal including G-TEO and the new G-MFCC.

2) Furthermore, depression detection was significantly improved with a new spectral roll-off range and improved further using a 2-stage mRMR and SVM optimized feature selection approach.

3) It was for the first time shown that adolescent depression could be detected from the acoustic speech parameters of parents. To our knowledge, it is the first objective evidence that there are clear emotional and acoustic differences between parents of depressed and non-depressed adolescents.
1.6 Thesis Structure

The remainder of this thesis is organized as follows:

Chapter 2 contains review of emotion recognition literature with an in depth review of speech production, physiological effect of emotions, acoustic features, current acoustic features studies and an outline of deep learning. This is followed by a literature review on current depression detection studies from an acoustic speech perspective. In addition a review of psychological validity of conversation modeling with respect to depression detection and possible conversation modeling approaches looks at applications of time series forecasting into prediction of emotions.

Chapter 3 outlines the data collection process in detail and explains the corresponding annotations used in the depression detection/analysis experiments. Likewise, the speech database used in the acoustic emotion recognition studies is outlined.

Chapter 4 explains the methodology for a new emotion recognition system based on 2D feature contours within a DNN compared to benchmarks feature vectors with Neural Network (NN), Support Vector Machine (SVM) and Gaussian Mixture Model (GMM). This is followed by the development of a new optimized multichannel DNN. Additionally, the formulation of the features, classification system, evaluation methods and experimental results are provided.
Chapter 5 describes the methodology of the conversation modeling system including background information on the Influence Model and new extensions. The feature extraction, classification, evaluation methods and experimental setups are outlined. The results are supplied for the EIM and depression classification. The Dynamic Emotional Influence Model (DEIM) and Higher Order Influence Model (HOEIM) are compared to the original first order EIM.

Chapter 6 outlines the methodology for prediction of emotional interactions in the parent-adolescent conversations using various types of random walks and autoregressive models. The evaluation methods and experimental setups are explained. Experimental results are presented and different methods are compared and examined.

Chapter 7 describes the methodology of depression detection based on acoustic speech parameters and machine learning. The evaluation methods, experimental setup and results are presented and discussed.

Chapter 8 provides a summary of research and results presented in Chapters 4, 5, 6 and 7. It compares results provided by depression recognition methods including approaches based on speech acoustics as well as conversation modeling parameters. This is followed by an overall conclusion and suggestions for future work.
Chapter Two:

LITERATURE REVIEW OF EMOTION RECOGNITION/PREDICTION, DEPRESSION RECOGNITION AND CONVERSATION MODELING SYSTEMS

2.1 Preview

The purpose of this chapter is to provide an overview of literature related to the four major research areas: (1) emotion recognition, (2) depression detection from speech, (3) depression detection from conversation model and (4) emotion sequence prediction. This literature review is a fundamental component to review and associate ideas between the research areas through the theory, methods and applications.

2.2 Literature Review on Emotion Recognition

The following subsections provide a literature review of studies related to emotion recognition research and outlines applications of emotion recognition. Concepts of speech production and the physiological links between speech and emotions are explained to provide rationale and justify the use of acoustic speech features for AER.

Past studies using acoustic speech descriptors to find correlations in emotions are reviewed and machine-learning classifiers and different features are compared. A detailed discussion on deep learning is supplied, including applications and in relation to emotion recognition and a brief outline of past multichannel studies.
2.2.1 Background of Emotion Recognition

Automatic speech recognition (ASR) and AER research have steadily increased linked with advancement of computing power and digital signal processing. From an engineering perspective AER aims to achieve efficient and accurate real-time systems. AER is widespread with growing collaboration between fields incorporated in modern technology (computers, smart-phones, cars) [510][511][512][513].

The presence and capability of computational devices is continuously increasing in our lives. Device functionality can be improved with AER to facilitate natural and realistic artificial intelligence (AI) [219][193]. AI combined with AER is required for user-friendly human-machine communications [261][221][233].

AER can enhance information to deduce the intended message, improve ASR rates and provide more realistic responses with emotion synthesis. AER has been used in AI for human-computer communication for adolescent-tutor tasks [402][403][192].

AER has been applied to call center operators and customers [257][258], car drivers [262][260], pilots under stress [228][229], interviews [264][263], parent-child conversations [265] and psychological diagnosis [236][264][109].

Emotions are expressed by changes associated with blood pressure, heart rate, facial expression and speech (linguistics and non-linguistic). Past AER studies have focused on audio [61][62][207], image [226][222][206], video [205][204], EEG [200][199], ECG [202] and other biological signals [203][201].

Audio analysis is appealing for AER as speech recordings are easily available and collected in a discrete, non-invasive manner. Audio data is simple and cheap to obtain, especially in real-time, from computers, phones or mobiles. Acoustic analysis has a large amount of literature available [60][61][62][64] with comprehensive reviews [431][347][233][172][190][191][207][405].
Speech can reveal information pertaining to health, age, personality, gender, accent, language, culture, socio-economic background and speaker state such as sleepiness, intoxication, mood, emotion and mental health [451].

Speech can be analyzed using linguistics that conveys meaning and semantic information used for ASR or language recognition (phonemes). Non-linguistic characteristics are useful in speaker or AER find characteristic regardless of language.

Evidence suggests ASR systems are adversely affected with emotional speech [228][229][230] and can be improved by integrating AER with ASR. Similarly, it is possible the performance of depression detection could degrade with emotional speech due to the link between depression and emotion [21][100][16][310][88]. Therefore, AER is an important concept in depression detection research.

### 2.2.2 Overview of speech production in relation to signal processing

Understanding physiological, biological and linguistic aspects of speech production is useful to determine meaningful acoustic features that correspond to emotional speech. There are four main stages in speech production defined as follows:

- **Respiratory (power (lungs, diaphragm)):** Air expelled from lungs up the trachea
- **Phonation (source (larynx, vocal cords)):** Air pressure opens glottis.
  - In voiced speech the air causes the vocal folds to vibrate.
  - Unvoiced due to turbulent airflow from constricted vocal tract
- **Oronasal (filter):** Air intake is filtered by the oral and nasal cavities
- **Articulation (filter (lips, teeth, tongue)):** Filtered by anatomical structures
During phonation physiological characteristics of speech are produced from periodic vibrations of vocal folds generated from airflow through the glottis opening and closing. The generated glottal pulse is an excitation source of voiced speech filtered by the vocal tract, cavities, and anatomical structures. A constricted vocal tract and a narrow airflow cause turbulent airflow resulting in unvoiced speech.

Speech production can be represented from a signal processing perspective by a linear system as the source-filter model, illustrated in Figure 2-1, that assumes the source (vocal folds) and filter (vocal tract) are independent used to shape the source.

The excitation generator produces a source, \( X(z) \), filtered by vocal tract, \( V(z) \), and radiation models, \( R(z) \) [338]. In voiced speech the source, \( X(z) \), is modeled by an impulse train, \( E(z) \), for the glottal pulse, \( G(z) \), to mimic glottis vocal folds vibration. Unvoiced speech the source is modeled as white noise, \( N_w(z) \), which is corresponds to turbulent air through a constricted vocal tract.
Vocal tract and radiation models are defined by transfer functions of the vocal tract, \( V(z) \), representing the shape of the vocal tract and the radiation transfer function, \( R(z) \), to simulate lips. A more detailed explanation in the production of speech and equivalent mathematical models can be found in [332][147][334][338].

The respiratory system structures controls air supply and affects the subglottal pressure and airflow rate and alters speech frequency and intensity. The phonatory system involves vibration of the vocal cords as air passes through the vocal tract creating a fundamental frequency and harmonics. The articulatory system then modifies the sound to create phonetic elements with different voice qualities.

2.2.3 Physiological rationale for acoustic AER from speech production

To supply rationale for acoustic AER, evidence is required of emotional correlation with physiological processes and hence acoustic properties. Evidence suggests respiration, phonation, and articulation are affected by emotions [158][321][339][17]. For centuries research has been conducted towards understanding physiological processes of speech production and the response to emotion [328][329].

During the 1960s research was conducted in relation to controls of laryngeal, pharyngeal and nasal structures to objectively measure emotion and stress in terms of prosody and electro-acoustic analysis [429][430].

In the 1970s and 80s Scherer et al. found emotion effects speech due to sympathetic and parasympathetic activation of the autonomic nervous system [321][17]. Emotions are neurophysiological related to the limbic system, which controls emotional expression and behavior [330][331].
In stressful and emotional conditions altered muscle tension affects vocal tract airflow [55]. Respiration affects sub glottal air pressure and vocal cord tension that changes, vocal cord vibration, vocal tract resonance, phonation and articulation [18] [339] and acoustic properties (i.e. frequency, harmonics and intensity) [18][17].

Williams and Stevens (1972) explained the sympathetic nervous system is aroused with particular emotions (anger, fear or joy), causing increased heart rate, increased blood pressure, drier mouth and muscle tremors resulting in louder, faster and strong high frequency energy [325]. Sadness has decrease in heart rate and blood pressure and more salivation with slower speech and minimal high frequency energy.

Physiological changes effect speech production evident in energy distribution across the frequency spectrum and duration of pauses of speech signal. These concepts have led to research into features to determine statistical correlation of emotions and machine learning for emotion recognition [73][60][61][62][322][323].

2.2.4 Existing Automatic Emotion recognition studies using acoustic features

Many studies have investigated acoustic parameters that relate to the physiological production of speech. Studies have compared various acoustic features, mainly low-level descriptors (LLD) (prosodic, spectral, cepstral and glottal) and statistical functionals (mean, variance, delta) extracted from speech for emotion recognition.

In the 1970s statistical properties of speech were examined and found correlation with emotion with fundamental frequency ($F0$), energy, speech spectrum, temporal parameters, articulation and glottal pulse information [325].

Nwe et al. provided a review of studies investigating properties of the “big six” (anger, joy, disgust, fear, surprise, sad) and concluded frequency, spectral energy, speech rate and voice quality parameters are reliable emotion indicators [356].
A) Prosodic

AER studies have investigated prosodic features such as $F0$, energy, zero-crossing rate, jitter and shimmer [62][333][237]. Under certain conditions prosodic features have outperformed other features [303]. A comprehensive feature review summarized anger was correlated with higher $F0$ [172] and more energy [234].

B) Spectral

Furthermore spectral features such as, FFT log energy [19], spectral energy in bands, flux, roll-off, centroid [62] are common in AER [233][191]. Individual spectral features and in combination with other feature categories have achieved good results in emotion recognition and even improved performance [61][359][358].

C) Cepstral

Cepstral features are based on the Inverse Fourier Transform (IFT) of the log-magnitude Fourier spectrum of a signal and can be portrayed as a complex, real, phase or power cepstrum [355]. Bogert et al. proposed the power cepstrum, which is especially effective in human speech analysis [350][351].

In the time domain the convolution of two signals, such as source and filter, are multiplied in the frequency domain and additive in the cepstral domain. Therefore, in the cepstral domain the vocal cord excitation (pitch) and vocal tract (formants) are separable which is useful for pitch estimation [350].

In the 1960s Noll and Schroeder proposed applying short-time cepstral analysis towards speech pitch estimation [348][349][350]. Rather than linearly spaced frequency bands Bridle and Brown suggested non-linear band-pass filters with logarithmic spacing after 1kHz and corresponding bandwidth increase [58].
For applications related to human speech Mermelstein suggested first transforming the spectrum into linear spaced frequency bands on the Mel-scale, to reflect the non-linear human hearing response [56][57][58]. The Mel-frequency cepstrum is a representation of short-term power spectrum and amplitudes that correspond to Mel Frequency Cepstral Coefficients (MFCC).

MFCC are widely utilized for speaker [353][59] and speech recognition [352][354] due to advantages of the cepstrum and characterization of the human auditory system. MFCC have been applied to AER [356][342][366][450] and in conjunction with prosodic, spectral or glottal features [61][62][73][225][226][360][174][342].

Nwe et al. investigated multiclass emotion recognition, of 6 emotions, using different cepstral parameters within a Hidden Markov Model (HMM) [356]. The results showed speaker dependent classification with Linear Prediction Cepstral Coefficients (LPCC) reached an average of 56.1% accuracy; MFCCs attained 59% and log-frequency power coefficients (LFPC) the best at 78.1%.

A study investigated AER from the FAU AIBO database for 5 emotion classes including: anger, emphatic, neutral, positive and rest. Speech parameters based on human perception modeled GMM with a weighted average accuracy of 41.5% using Perceptual linear prediction coefficients (PLP) and MFCC at 45.3% [366].

**D) TEO**

Recently Teager Energy Operators (TEO) have become popular providing insights into non-linear speech models, to represent vortex flows [68][69][70]. TEO have been successful in stress and emotion recognition [235][189][64][19][55][263][228].

Emotion recognition was investigated for five emotions using a TEO based feature set comparing PNN and GMM classifiers. The highest multiclass accuracy was 62% using the GMM, which consistency outperformed the PNN [235].
E) MFCC and TEO

Nwe et al. investigated both stress and emotion recognition using acoustic features including MFCC, LPCC and log-frequency power coefficients (LFPC) [19]. The study proposed extracting the LFPC from the TEO in both time (NTD-LFPC) and frequency (NFD-LFPC) domain. Results indicated, for 6-class AER, NFD-LFPC and NTD-LFPC were optimum at 86% and 79%, MFCC lower at 77% and LPCC at 67%.

A study investigated MFCC and TEO based MFCC (TEMFCC) in noisy environments with a range of SNR levels of additive white and pink noise to the EMO-DB. 7-class speaker independent GMMs achieved an average AER accuracy of 46% and 42% for white noise using TEMFCC and MFCC and for pink noise 45% and 47% [189]. At low noise and clean conditions MFCC are optimal and TEMFCC is more robust to noise, outperforming the MFCC accuracy as the noise level increases.

MFCC are a commonly used cepstral feature utilized for emotion recognition but are easily affected by noise that is insensitive by human perception [363]. TEOs have largely eliminated the effect of noise [364] and the degradation of MFCCs has been improved using the TEO to extract new cepstral coefficients [363][365][189].

F) Glottal

Glottal waveform and related features are used less often, compared to the speech waveform features, with limited studies exploring glottal waveform features. This in part could be due to the difficulty in extracting the glottal waveform features with estimation methods not robust given a noisy data source [515].

Speech recognition systems can be improved by combining the glottal characteristics with acoustic parameters [168][63]. Studies have shown glottal features provide a strong correlation with speech characteristics[195] and speaker identification [208][388][389].
The glottal waveform is affected by emotion and extracting features from the glottal waveform can be beneficial. Emotional speech has been shown to greatly affect the glottal waveform [401]. Glottal waveform parameters have been considered in stress [194] and emotion recognition applications [194][168][197][391][63][64]. Glottal features are quite robust with severely low-pass filtered speech and additive white Gaussian noise compared to other features [392].

Past studies have successfully improved speech related classification tasks by deriving TEO parameters from the glottal waveform [238][65][147]. In addition, MFCC derived from the glottal waveform have been investigated in relation speech recognition tasks and not always improving upon regular MFCC[342][367].

Iliev et al. presented glottal waveform derived MFCC for AER and compared to speech waveform MFCCs. The study proposed multiclass AER (angry, neutral, happy and sad) using optimum-path forest classification (OPF), GMM, SVM, NN, k-NN, Bayesian classifier (BC) and C4.5 decision tree. For all classifiers and MFCC orders (4,6,8,10 and 12) the glottal waveform improved recognition rates [342].

In 2009 ICASSP, Sun et al revealed prosody is not enough to discriminate between certain emotions. The work showed glottal parameters outperformed the accuracy of prosodic features. Misclassified emotions using prosodic features were significantly improved using glottal ratio parameters [63].

Sun et al. compared TEO_CB_Auto_Env parameters, time-based glottal parameters, spectral and prosodic features using SVM training [64]. TEO and glottal parameters outperformed prosodic and spectral features. TEOs were stronger, compared to glottal, in classifying binary states of activation and power dimensions and glottal parameters were best classifying binary states of expectation and valance.
G) Multiple Features

Vogt et al. used the EMO-DB for simultaneous classification of seven emotions using a Naïve Bayes classifier [360]. A combined feature set including pitch, energy, MFCC, pauses, duration and speaking rate (totaling 1280 features) achieved 69%. A correlation-based feature selection method improved the accuracy to 77.4%.

Schuller et al. explored a two-stage model to provide temporal resolution by chunking speech-turns and extracting features: $F0$, energy, formants, shimmer, jitter and MFCC from EMO-DB [175]. An SVM attained 46.7% with all features, 51.4% with feature selection and 70.6% with turn-level mapping and feature selection.

In 2009, Schuller et al. provided a benchmark of nine databases under equal conditions with standardized 4 clustered classes on arousal and valance levels [62]. Zero-crossing rate (ZCR), log energy ($LogE$), $F0$ and MFCC were extracted from the EMO-DB and reached 84.6% (HMM-GMM) and 73.2% (SVM) for 4-class AER.

Vlasenko et al. investigated emotion recognition with a large feature set of LLD comprised of 1407 features and reduced to just 76 with optimized feature selection. Tests on the EMO-DB attained an accuracy of 89.9% by combing turn-level and frame-level classification [516]. A downside to the study was the small proportion of the entire feature set used in the optimal set that increased computational cost with many redundant features.

Schuller et al. investigated 276 statistical hi-level prosodic, articulatory and speech quality features and applied optimal feature selection for emotion recognition from EMO-DB [517]. The entire feature set of achieved 84.8% with minimal improvement using feature selection reaching 87.5% in SVM. An optimal segmentation scheme and fusion of segments with respect to classification and combination of diverse timing levels achieved a result of 96.5%.
H) Summary Features

Many of the acoustic speech feature parameters used in speech recognition tasks are based on Low-Level Descriptors (LLD) with studies primarily focused on determining the ideal acoustic parameters. Recent studies utilize excessively large feature sets such as the well-known research with The Munich open-source large-scale multimedia feature extractor (openSMILE) [533][535] or The Geneva minimalistic acoustic parameter set (GeMAPS) [534].

Studies have explored optimising a set of feature parameters, which can differ on database, conditions and setup. Research has concentrated on capturing all characteristics with a high-dimensional feature set but are prone to over-fitting, poor generalization and are computationally expensive.

Rather than a “brute force” of investigating an excessively large range of acoustic features an alternative is to choose a set of features that have shown promising results individually that can be examine with augmentation, modifiers and as new representations.

Mel-frequency cepstral coefficients [356][342][366][450] and Teager Energy Operators [235][189][64][55][228] are an appropriate choice as the baseline feature sets, as they have been effective and successful in many AER applications. MFCC and TEO have improved performance derived from the glottal waveform (augmented) [238][65][147][342][367], shows improvement with derivatives (modified) [138][518] and as a vector could be reformed as a 2D image (representation).

Instead of re-working an in depth analysis of all possible features and combinations a restricted baseline set is explored and expanded and concentration of ideal modeling and classification methods can be conducted
2.2.5 Deep Machine Learning and classification

Machine learning is an exciting field to learn models to classify unknown samples and predominantly focused on single-channel classification systems. The majority of existing work investigated classification systems using Support Vector Machine (SVM), Gaussian Mixture Model (GMM), Neural Networks (NN), K-nearest neighbors (KNN), Hidden Markov Models (HMM) and hybrids.

Recent advancement in deep learning techniques has led to a resurgence in unsupervised learning, supervised learning and in particular Deep Neural Networks (DNN). A DNN has multiple hidden layers that represent different aspects of features at each layer and have ability to encode higher order network dependencies [170][89].

A broad review of deep machine learning techniques was explored by comparing convolutional neural networks, deep belief networks and variations [278]. Tasks using shallow neural networks have performed worse compared to the DNN, due to the ability of deep nets to build up a complex hierarchy of concepts. [170]

2.2.5.1 Deep Neural Network (DNN) Applications

Deep learning has proven to be powerful classification approach in many applications including handwritten character/digit recognition [114][75][71], face recognition [223], object recognition [279] and motion capture modeling [215]. Image recognition applications use the actual image to be trained as feature detectors using a DBN [72].

Deep learning models have been used in acoustic modeling for audio and speech recognition systems [224][187][183][186][196][93]. These applications use images of time-frequency speech spectrograms, a direct image of the speech waveform or secondary features from LLD.
Using TIMIT corpus for phoneme recognition the speech segmented into 25-ms Hamming windows with 10-ms overlap and 12th-order MFCCs, energy, and the first and second derivatives were extracted. The DBN managed to outperform other techniques with an optimal phone error rate of 23.0% [187]. Compared to other studies such as a recurrent NN at 26.1% [217], or a GMM reaching 33% [218], and the previous benchmark of 24.4% using a heterogeneous classifier [216].

2.2.5.2 Deep Learning for emotion recognition

Many deep learning methods including: DNN, Convolutional Neural Network (CNN), restricted Boltzmann Machine (RBM) and deep belief network (DBN), have recently been used towards emotion recognition [227][225][73][226][404][536]. These systems have demonstrated applicability to learning useful structures from complicated, high-dimensional data to model and generate a large range of facial expressions from sparsely labeled data in unsupervised learning [74]. Many studies have been implemented using images of faces or directly from the raw speech.

A study used deep generative models for multiclass AER with using facial images of seven emotions [188]. Raw pixels at the DBN achieved 90.1% accuracy while a linear classifier only achieved 83.8% and a Gaussian SVM 85.1%. Another study investigated AER by combining CNN with Long Short-Term Memory (LSTM) networks to automatically learn the best representation of the speech signal from raw representations [536].

Instead of raw time or frequency domain features it is possible to learn using acoustic features. A study used spectrograms generated with principal component analysis (PCA) and whitening to create unlabeled data patches for RBM unsupervised learning to generate features from the EMO-DB. A second stage trained a SVM reached 71% using the RBM features and the raw features only reached 23% [227].
A study proposed multimodal DBN models with acoustic (prosodic, spectral, and MFCC features) and video (Facial Animation Parameters) features for multiclass AER using the IEMOCAP database (anger, happy, neutral and sad). The DBN outperformed the baseline classifiers with a 3-layer DBN attaining 73.8% with prototypical data and a 2-layer DBN 70.5% [226].

A study proposed using DNN to learn from LLD features including: zero-crossing rate, log-energy, fundamental frequency, centroid, flux, roll off and MFCC. Generalized discriminant analysis (GerDA) extracted parameters from DNN feature detectors. Valence and arousal levels of emotions were classified using Mahalanobis minimum-distance measure and outperformed a SVM [73].

Le and Provost proposed and evaluated a hybrid a DBN-HMM using MFCCs, energy and the first and second derivatives creating a 39-dimensional vector. Five-class AER achieved a top result of 46.35% average recall [225].

2.2.5.3 Multichannel Classification

Following on from single channel AER, containing a single feature set, this section investigates using multiple features. Features that perform poorly or well individually be improved with additional features by pooling strengths and providing mutually helpful information.

This can be achieved though early integration (feature fusion), which concatenates features (merged/fused) into a matrix. An alternative approach is late-integration using independent feature models fused at a decision or score level with multichannel classifiers such as SVM [421][163], GMM [421][92], GMM-HMM [91][90], SVM-GMM [421], MLP-GMM [421], CNN [198] and DNN[93].
A study compared feature fusion and late-integration (decision and score) and found that feature fusion was weaker than late-integration [421]. Moreover early-integration doesn’t take the lack of synchronicity between features into account [90].

A 2016 study examined decision fusion of multimodal features in an extreme learning machine (ELM) with multiple databases. The study investigated seven-class AER and reached a top result of 50% using multimode weighted decision fusion. The weights were determined from a pool of 50,000 random weight matrices [523].

Schuller et al. used early fusion of acoustic and linguistic features in multichannel AER using the EMO-DB [60]. Acoustic features included ZCR, $F0$, formants, energy, harmonics-to-noise-ratio, MFCC and FFT. The linguistic features were extracted using bag-of-words. The top 75 features in an SVM reached 87.5% average accuracy and with integration of linguistic cues improved by 3.51%.

2.2.6 Summary Emotion Recognition Literature

This section supplied rationale of using speech for AER with examination of acoustic features showing the discrimination ability of glottal features, TEO and MFCC are encouraging. Most AER studies have used SVM, GMM, KNN, HMM and NN with a recent resurgence of Deep Learning (DNN) in practical applications.

This research investigates the effectiveness of DNN for pair-wise (binary) and multiclass AER. The novel aspects are the DNN inputs given as higher-level features representing 2D black and white images depicting contours of acoustic parameters.

Weighted multichannel classification improves performance of baseline classifiers and can be applied to the DNN. The weights could be proportional to single classifier scores [92][421], determined with unsupervised learning [90], exhaustive searches [523], grid-searches [421] or in this case with supervised learning and a global optimization algorithm as a new optimized multichannel DNN (OMC-DNN).
2.3 Literature on Depression Recognition from Speech

Depression is a serious disorder and has directed psychology research and recently considerable interest from engineering viewpoint to investigate automatic depression detection. The following subsections provide a literature review related to speech depression detection from psychological, engineering and collaborative perspectives. Rationale for depression detection from speech is explained by physiological interoperations and link to acoustic features. Studies analyzing acoustic correlates to depression and automatic speech depression detection applications are discussed.

2.3.1 Background of Automatic Depression Detection Modes

Recently automatic depression detection has become a popular trend from an engineering perspective, with development aided by advancements of computers, digital signal processing and medical equipment. Depression is an affect (emotion) regulation disorder [24], associated with emotional disturbances such as excessive sadness and reduction in expression emotions [449]. This has led to advancements in depression detection through replicating affective computing method (AER).

Biological markers have sparsely been examined in depression with a correlation with low serotonin levels in depression, but no definitive characteristics [479]. Correlation of brain structure from medical imagery has been explored in depression [455]. Electroencephalogram (EEG) signal analysis has effectively distinguished depression [412]. Other studies have analyzed ECG signals [468][469] and blood pressure[468] in depressed patients finding correlations with treatment.

Biological signal analysis is difficult to validate as data and modeling can require elaborate, costly, invasive, time-consuming and computationally expensive techniques, leading to problems in a real world application. This could be why there are limited studies using biological signals in depression detection.
Face tracking developments have made it possible to apply facial analysis towards depression from image/video. This can be from a facial expression perspective using the entire face or sections, with facial action coding systems (FACS) using manually annotated action units (AU) [443][445][486], active appearance models (AAM) [458][443][418][417], eigenface and fisherface features [454] or Gabor wavelet facial landmarks features [456]. In addition eye movement/gaze [418][415], head pose/movement analysis [486][418][416] and motion history histograms [487] have been used in depression detection.

Speech analysis is beneficial for mental health monitoring [530][531], with readily available databases and natural non-invasive collection for practical application via microphones. With modern technology this is becoming easier to obtain including from computers and telecommunications (hotlines, phones and mobile apps). Although depression detection acoustic speech analysis is a relatively new area of research it is facilitated by AER, with decades of active research.

As discussed in Section 2.2 speech analysis using linguistic and non-linguistic cues has provided strong physiological links in emotion [17]. Studies suggest a strong correlation between expressed emotions, mood changes, and stress with clinical depression [21][100][16]. Considering the proficiency of speech analysis in emotion recognition and the correlation between emotion and depression it would be expected the concept could be applied towards detect clinical depression.

Studies have shown potential in depression detection using linguistic analysis, such as bag-of-words with codebooks, but only in multimodal classification along with acoustics/images [458][163]. Moreover, linguistic analysis depends on language and the goal is to detect depression regardless of linguistic information or language. Subsequently a non-linguistic acoustic speech analysis approach is idyllic.
2.3.2 Physiological Correlation of Vocal Indicators of Depression

For over a century it has been recognized that clinical depression is associated with changes in speech [395][396]. It has been long recognized that speech and speech patterns contain information regarding psychological state (depression) [293]. Psychiatric textbooks indicate diagnostic value of examining speech deviations in affective disorders, particularly depression [393].

In major depressive disorder (MDD) neurophysiological changes tend to alter motor control with psychomotor retardation [474][475]. Biological markers of depression have been found with motor function disturbances exhibited by slower movements in muscle activity [393]. Speech is generated by the neuromuscular system and is affected by depression in quantifiable parameters [394][477].

The physiological fluctuations lead to alteration in speech production by shaping the vocal folds (source) and vocal tract (filtering). Recent studies have suggested airflow in speech production show distinguishable physical manifestations in depressed subjects and have significant effect on vocal folds [462][463].

Changes to vocal tract properties possibly caused by amplified vocal tract tension [432][109], a lack of motor coordination [488][477][482] or medications drying out the vocal tract [109], can alter speech spectral properties [482]. It has been theorized a person suffering depression requires more articulatory effort to produce and maintain speech [437][432]. This effort causes prosodic, articulatory and phonetic changes in depressed speech [477][480][481][482].

This could explain why vocal characteristics change with mental condition [437][109][477][478][457]. Speech in depressed subjects is discernible and characterized as dull, monotonous and lifeless by clinical experts [474][293]. This reflects the quality perception including monotony, slur, hoarseness, and breathiness.
These qualities can be related to acoustic features, for example monotony can be associated with prosodic features (\(F0\) modulation, energy, speech rate) [457]. Likewise hoarseness and breathiness are related to prosody (jitter, shimmer) and slur is connected to vocal tract articulators (formants) [457]. Specifically, a symptom of depression is psychomotor retardation, which can tighten of the vocal tract that tends to affect formant frequencies [109][432].

2.3.3 Vocal Indicators of depression

Research has identified objective vocal parameters that are relevant to physiological processes. Acoustic analysis has been applied towards many mental health problems including schizophrenia [439][156], stress [375], affective disorders [17][18][376], depression [156][397][440][441][376][477][529] and psychopathy [442] observing changes in speech prosody. Early studies investigating aspects of speech that correlate to depression were generally conducted from a psychology or psychiatric approach.

In 1930 a psychiatrist, Zwirner, began to analyze objective acoustic parameters. The fundamental frequency (\(F0\)), tempo and pause time of speech were measured with a device to track speech variability in depression[371]. The pioneering technical method led to many subsequent studies. Continued investigations during the 1930s to 1960s found depressed patients had a slower speaking rate (tempo), increased pause time and lower \(F0\) with less variability. [372][373][374].

A subjective assessment by Darby and Hollien (1977) found listeners perceived significant changes in prosodic speech characteristics (\(F0\), loudness, speaking rate and articulation) in depressed patients before and after treatment [104]. Another study by Darby et al. (1984) subjectively analyzed speech of 13 depressed patients before and after treatment. It was confirmed depressed patients exhibited lower \(F0\), less \(F0\) variation and lack expression (monopitch and monoloudness) [397].
Late 1980s studies by Nilsonne objectively analyzed speech parameters during clinical interviews of 16 depressed patients, during a depressive episode and after clinical improvement. $F0$ standard deviation, the rate of change of $F0$ standard deviation and $F0$ average speed change were correlated with higher levels after depression recovery [98]. Similarly, follow-up studies with 28 subjects found a reduction of $F0$ and pause time was correlated with mental state [419][436]. Kuny et al. analyzed speech of 30 depressed patients and found a strong correlation of $F0$ related parameters ($F0$-contour, $F0$-bandwidth and $F0$-amplitude) [99].

In 1965, mood assessments of 32 depressive patients were paired with recorded interviews. Large adjustments in mood were detected from spectral features such as an upshift of the first formant and a power increase in higher formants as patients became less depressed [390]. Contradictory studies found noticeable decrease in the 2nd formant for depressed compared to controls [432] and an increase (decrease) in spectral energy in the lower (higher) frequency bands after therapy [156].

Scherer et al. analyzed a range of acoustics relating to prosody including $F0$, formants, intensity, speech rate, energy and parameter variabilities. Significant correlation was found in vocal indicators with affect disorders such as anxiety, depression and stress [17][18][375][376]. Evidence suggested sadness and depression are associated with a decrease in energy and loudness as well as a rise in jitter[376]

Recent studies examined vocal parameters in correlation to depression severity based on scale ratings (HAM-D and QIDS). Negative correlation was found in energy (reduced energy with increase depression severity), speech rate (slower rate) and formants (down shifted 1st and 3rd formants) and $F0$ had positive correlation (lower $F0$), [457][437]. It was determined patients responding to treatment had significantly larger $F0$ variability, shorter and less pauses and spoke faster [437].
To summarize, although findings are not entirely consistent there are many agreeable observations. Most studies determined similar correlations of features and depression with minimal contradictions as follows:

Subjective [104][397] and objective [371]-[374][436][419] studies found $F0$ and variability are lower in depressed. $F0$ is negatively correlated with depression severity and decreases as a patient gets worse [457][437] and increases after treatment [438][98]. The lower $F0$ range indicates monotone speech, commonly observed in depressed patients [293], and a lack of expression indicated by less $F0$ variance [397].

Formants are significantly distinguishable in speech of depressed subjects compared to healthy people [156]. Generally the presence of depression is correlated with lower formant frequencies [437][432]. This is due psychomotor retardation, a symptom of depression, which tightens the vocal tract alters formants[109] [432].

Depression is associated with decreased loudness and energy [104][376] and further reduced with increased severity [457][437]. After treatment a power increase in higher formants has been observed [390]. This was contradicted with a spectral energy increase (decrease) in lower (higher) bands observed after therapy [156].

Depressed patients present a slower speaking rate and increased pause time in subjective [104] and objective studies [371]-[374][100][419][436]. Pause time reduces after treatment [398]. Decreased speech rate is correlated to depression severity [457][437].

Many early depression studies have subjectively and objectively evaluated prosodic and spectral speech parameters as physiological indicators of depression, severity and treatment efficacy [437][477][529]. Differences in $F0$, loudness, speaking rate and articulation have led to increasingly complex acoustic parameters and sophisticated learning systems [437].
2.3.4 Acoustic speech parameters in machine learning Depression Detection

Studies conducted by psychologists are generally concerned with overall patterns of speech from statistical prosodic speech measures. In contrast recent engineering approaches have utilized frame-by-frame acoustic features extracted from speech. There has been a growing interest in affective computing and signal processing fields in developing automatic systems, demonstrating efficient acoustic analysis systems.

It is popular to train a machine-learning algorithm to classify depression based on acoustic features known as LLD and statistical functionals. The most significant studies in depression detection from speech, concentrate on feature categories including: prosodic, spectral, glottal, cepstral (MFCC) and Teager Energy Operators.

France et al. studied long-term acoustic properties in correlation to depression severity and suicide risk from 42 subjects [158][109]. It was determined $F0$ was ineffective in distinguishing depression where as formants and power spectral density (PSD) were superior discriminators in depression levels up to 94%.

Ozdas et al. examined acoustic speech parameters from 30 subjects (10 suicidal, 10 depression and 10 controls) determining glottal slope showed statistical significance and no significance using jitter. Jitter and glottal slope features trained multivariate maximum likelihood classifiers to discriminate suicidal and control at 85%, depressed and control at 90% and depression and suicidal at 75% [106]

Moore et al. used a clinical database of 15 depressed and 18 controls and compared prosodic and spectral features. Speaking rate was the weakest feature, prosodic features in general performed worse than spectral [139]. A follow-up study used quadratic discriminant analysis to select acoustic features (prosodic, spectral and glottal) with the highest depression classification. Glottal features reached 87% (males) and 94% (females) and combination was 91% (male) and 96% (female) [105].
A study compared acoustic features and facial image analysis in depressed patients receiving clinical treatment. Facial actions features trained an SVM reached 88% for FACS and 79% for AAM. A logistic regression classifier with prosodic features predicted response to treatment at 79% accuracy using $F_0$ [417].

The aforementioned studies have achieved encouraging results in depression classification, although the datasets are quite small and may not be meaningful for clinical applications. Furthermore the features are limited, generally simple one-dimensional parameters. Later studies have employed more sophisticated acoustic features commonly used in speech analysis tasks (emotion, stress, speaker characterization).

Low et al. used a large clinical database, Oregon research Institute database (ORI), with 139 parent-adolescent interactions (68 depressed and 71 controls) [138]. The study proposed MFCC, delta MFCC and delta-delta-MFCC to detect depression in adolescents using a GMM. Incorporation of both delta and delta-delta of MFCC improved classification by 3%. Gender dependent models (GDM) outperformed the gender independent model by 8% with 58% for males and 60% for females.

Low et al. continued on from their preliminary study, using the ORI-DB, with MFCC, $LogE$ and TEO with GMMs for content dependent depression detection. The GDM improved accuracy with TEO attaining 54% for males and 61% for females; MFCC+$LogE$ reached 65% for both males and females [107].

A study using the ORI-DB extracted LLD with prosodic ($F_0$, formants, $LogE$, jitter, shimmer), spectral (spectral flux, centroid, entropy and roll-off), MFCC and TEO features and used separately and in combinations in GMM. A combination of TEO, $F_0$, $LogE$, shimmer, spectral flux and spectral roll-off gave the best result of 77.8% for males, the top result for the females was 74.7% using only TEO [108].
A follow-up study, with the ORI-DB, explored depression from four acoustic categories: Prosodic, Spectral, Cepstral, TEO and Glottal. A GMM with P, S and G features reached 67%, 69% and 73% for females and 60%, 61% and 75% for males. A combination of G+P+S improved upon S+P alone (69% males and 75% females). Overall TEO was the best with 79% for females and 87% for males [140].

A study investigated spectral, prosodic and MFCC features with GMMs from a subset of read speech in the Black Dog Institute dataset and concluded MFCC were the best feature in speaker dependent and independent depression detection[434].

A pilot study in prediction of depression in adolescents proved it is possible to identify those likely to develop depression within 2 years [65]. The examination was limited to a small database of 30 participants (15 at risk and 15 controls) using the ORYGEN corpus. Glottal features were the strongest predictors of depression at 69% and the TEO only 52%. TEOs derived from the glottal waveform and a combination of glottal and TEO features reached 65% and 63%, improving the stand alone TEO.

Another study confirmed speech features, from four acoustic categories, prosodic, glottal, spectral and TEO, could predict early signs of depression in adolescents. A multichannel GMM, with all features, achieved 73% for person-dependent and 61% in utterance approach [92].

A study compared spontaneous and read speech, from the Black Dog Institute dataset, for depression detection using LLD including $F_0$, $LogE$, voice quality, MFCC, HNR, formants, energy, shimmer, jitter, intensity, loudness. Individual features used to train SVM models ranged from 63% for $F_0$ and 78% for $LogE$ and determined spontaneous speech was better compared to read speech [422].

43
Alghowinem et al. provided a comparative examination of classifiers (GMM, SVM, MLP and Hierarchical Fuzzy Signature (HFS)) in depression detection, from a subset of the Black Dog Institute. Loudness achieved 85%, MFCC 69.4% and overall top of 91.7% using all features with SVM-GMM [421].

Recently, research has shown the importance of acoustics in detecting depression severity. A study examined prediction of self-reported depression severity level, indicated by the Beck Depression Index (BDI), using the AVEC database [465][453]). LLD were extracted and transformed into iVectors that trained support vector regressors to model depression severity level. RASTA-PLP, MFCC and spectral outperformed the prediction performance of prosodic features [424].

Mitra et al. investigated HAM-D depression score prediction from 17 patients during admission, follow-ups and release (VU-PTU dataset). Significant acoustic differences were apparent in varying levels of depression severity (HAM-D) and predicted depression severity from articulation, phonetics, spectral, prosodic and perception features and showed capabilities in cross-corpus prediction [426][427].

To summarize, acoustic depression detection has concentrated on acoustic categories including: prosodic, spectral, MFCC, glottal and TEO. Studies have found spectral features generally outperform prosodic in depression classification studies [158][139][108][140][424][109].

TEO has been used in depression detection [147][107][65], and in certain cases outperformed baseline features and combinations [108][140]. MFCC are widely used in detection [108][138][140][427][421][422] and has outperformed spectral, prosodic and/or TEO in some cases[107][424][434][435].
The important role of glottal waveforms on clinical depression and the effectiveness in depression recognition is well documented [140][92]. Glottal features have been strongest compared to benchmark features in many studies [106][65][105]. It is suggested this was due to the glottal domain representing the source of voice production and its sensitivity of subtle voice changes.

Studies have consistently shown the critical role of glottal features, which added to standalone feature categories, improves depression detection (i.e. P+G, P+S,P+S+G) [140][105]. Deriving TEO from the glottal waveform has improved depression classification [65][147].

2.3.5 Limitation with acoustic depression recognition studies

In general, studies concerned with finding acoustic correlations with depression from a machine learning approach have shown consistencies, generally agreeing on the effectiveness of acoustic features, but there are discrepancies. Much of contradictions, related to ideal features and classification setups, can be put down to the complication of validating studies with dis-similar and limited databases.

Although depression detection systems have increased in the last few years there is still considerable research to be conducted. Much of the work is limited to due database restrictions with many not large enough to gather relevant and statistically valid results. Currently work has been moving towards larger databases and acoustic cues of depression in adolescents within spontaneous speech interviews.

The studies that used conversational databases were limited to analyzing the speech of the individual of interest [140][108][107]. The speech acoustics of their parents is not utilized and therefore the potential of the database has not been reached.
It can be noted that as of yet no studies have published results on the possibility of detecting depression in adolescents from their parent’s speech. This could be an important concept, considering ethics of studies related to children; instead the parent can be recorded for a screening of his or her own children.

This idea has already been explored in relation to children with autism in interactions with psychologists. A study demonstrated that the child’s prosodic features correlated with autism severity and introduced the concept of modeling a psychologist’s prosodic behavior to indicate the child’s ratings [520]. It was found that the psychologist’s features were more predictive than the child’s acoustics.

Other studies investigated autism detection and severity levels through acoustic analysis of spectral, prosodic and cepstral features [519][521] as well as turn-taking features [522] of both the child and psychologist.

The interoperation was that the psychologists adjusted their behavior based on the child's social-communicative impairments [521] and that psychologist attuned to the child’s behavior needs deliberately or spontaneously [520][519].

This could be worth investigating with family relationships bearing in mind psychological studies that have shown a correlation between parent-adolescent interaction and depression [24][135][136][137][144][145][157][319][134][320].

This gap in knowledge is addressed in this thesis with the proposed introduction of adolescent depression detection from the acoustics of their parents towards adolescent detection.

Another fundamental disadvantage of using acoustic analysis is the reliance on recording quality such as the environment, room acoustics, microphone position, microphone quality and SNR. Also potential mismatches between training and testing data with language, gender and age dependencies an issue in cross-corpus tests.
Moreover, acoustic analysis does not consider emotions expressed through speech and face, body and gestures can help determine signs in depressive symptoms by psychologists. The interest in human behavioral and conversational analysis has increased particularly to interpret social signals for mental health applications.

2.3.6 Summary Speech Depression Recognition Literature

This section outlined the reasoning of using speech for depression detection with examination of acoustic features showing the discrimination ability of spectral, glottal, cepstral and TEO-based features are promising. This research investigates the effectiveness of adolescent depression detection using a range features (prosodic, spectral, cepstral, TEO and glottal). The important role of glottal waveforms on depression and effectiveness in depression recognition [140][92][106][65][105] has led this research to introduce the glottal waveform extracted MFCC.

In general acoustic depression detection studies have been restricted to the analysis of the subject of interest. A novel aspect of this research is in relation to the hypothesis that it could be possible to detect depression from another speaker. This concept is addressed by examining the effectiveness of adolescent depression detection using acoustic features extracted from parent’s speech.

Limitations of acoustic analysis are addressed in this thesis with a new conversational modeling system for depression detection that does not rely or depend on recorded speech signals of any speaker. Instead depression detection is based on characteristics and correlations of emotional sequences, of the parent and adolescent, manually labeled from observed audio and visual cues.
2.4 Literature Review for Depression Recognition using Conversation Analysis

The aim of this section is to deliver justification of the psychological validity of conversation modeling for depression detection. This achieved through a review of psychology literature in relation to depression and the relation to family interactions. An outline selected conversation models are described including possible generative models capable of being used as a conversation modeling system.

2.4.1 Psychological validity of depression diagnosis from conversations

Conversation analysis is an important field of research from both an engineering and psychological or social sciences perspective. An understanding of conversational interaction patterns can useful for behavioral analysis with detection of conflict and dominant personalities, determination of roles within conversations and mental state.

Those with depression generally express emotions such as sadness and anger more than those without depression [445] with expressed emotions more evident is those that suffer from depression [310][311]. Emotions play an important role in conversations, especially during parent-children interactions [24][135][136].

The interaction of the emotions and the participants of conversations could have discriminating factors for depressed and non-depressed speakers. Existing research shows strong evidence to support that family interactions have a connection to depression in adolescents [136][134][137].

During conversational interaction each participant influences each other’s behaviors [341]. There is vast research in social sciences aimed at understanding interactions between individuals and factors that influence their behaviors. In an early psychological study, interactions between speakers were explored in relation to confidence levels influences from others [28].
Psychological researchers have extensively studied family relations in attempts to reveal how the risk factors for depression relate to parent-child conversations [24][134][135][136][137][145][157][320][144][319]. Many studies in child psychology indicate that family relationships and interactions are critical factors in the development of depressive symptoms.

Sheeber et al. suggested the quality of family relationships/interactions is important to understand symptoms of adolescent depression [135][136] [137]. Studies shave shown depressed children are more likely to experience an adverse family environment [145][157]. This is characterized by elevated levels of harsh and conflictual interactions and an absence of support and warmth [136].

Moreover adverse family environments are associated with depressive symptoms of adolescents and difficulties in regulation of negative emotions [137] [24]. Negative reciprocal parent-children process reinforces the children’s depressive behaviors [137][135][144].

A study, using a database collected by the Oregon Research Institute (ORI), examined family relationships between 233 adolescents. The participants either suffered from a major depressive disorder or were controls. During interactions it was revealed that depressed adolescents struggled to identify the mood of their parents and often misinterpreted their emotions, known as “mood bias”. The depressive state of the child was revealed to be directly related to the amount of bias in mood [320].

A similar study found depressed adolescents have the tendency to misattribute happy and sad facial expressions as anger. Furthermore adolescents with depression incorrectly interpret neutral as anger [319]. In particular depressed adolescent’s over-perceived aggressive behavior and under-perceived positive affect, potentially lead to
elicit adverse reactions from others [320]. These findings help understand processes that contribute to adverse relationships with depressed adolescents.

Based on the evidence presented, emotions and family interaction play a role in depression. This provides rationale to create a conversation modeling system that can extract information about emotional influences in parent-adolescent conversations. This could provide an initial diagnosis and additional information to facilitate therapy and correct maladaptive emotional interactions.

2.4.2 Studies using psychological models of conversations for depression

Contributions made by psychology and social sciences are predominantly concerned with subjective measures describing the conversational behavior in qualitative rather than objectively measurable parameters [144][180].

Conversational interactions have been of keen interest to social and health sciences, linguistics, and behavioral psychology, as well as signal process engineering. Recently, collaboration between disciplines has attempted to better understand the conversational and emotional structures using statistical models.

A widely used statistical model for conversation analysis is the Actor Partner Independence Model (APIM) proposed by Kenny et al. [181][182]. The APIM models dyadic relationships and conceptualizes inter- and intra-dependence based on subjective observation scores from predictor variables describing emotion, cognition and behavior.

APIM extensions have applied trajectories to predict time related changes characterizing speakers [298][299][180][300]. An application of APIM from 203 mother–adolescent dyads showed interpersonal influence on attachment security is bidirectional with degree of the mother’s influence diminished with age [297].
Kuppens et al. proposed a concept of emotional inertia (speaker dependence) to characterize psychological maladjustments of resistance to emotion changes in conversations [88]. The study investigated adolescent depression using ORI-DB, 141 adolescents (72 depressed and 69 control), with LIFE annotations converted to binary behavioral codes and used in three-level logistic autocorrelation regression models. Depressed participants had significantly higher autocorrelation for happy, angry, and dysphoric emotions. This represented depressed individuals having high emotional inertia, resistance to external influences and preservation of current emotional state.

Kuppens et al. followed with another study using behavioral codes of 165 adolescent-parent conversations to predict depression after 2.5 years. Results indicated emotional inertia (self-dependence) is highly correlated with depression. Adolescents at risk of depression have higher emotional inertia for negative and positive emotions [87].

2.4.3 Studies using generative and probability models for social analysis

Concepts for studying interactions has been achieved in many fields including psychology, sociology, political science, economics, biology, geosciences, and engineering [27].

Psychological studies have determined depression risk factors are based on conversational emotion interaction. The psychological perspective is concerned with subjective/qualitative measures that describe speaker behavior.

From an engineering perspective, quantitative results can be generated more complex mathematical, statistical, generative or probabilistic models. Various models could be appropriate to analyze speaker interaction such as Hidden Markov Models (HMM), Dynamic Bayes nets (DB-N) and Influence Model (IM)
A) Hidden Markov Model (HMM)
Social analysis has been analyzed with HMM including: multi-stream [11], coupled [283][284][15][10], layered [13][295], linked [103] and asynchronous [12]. McCowan et al. showed coupled and layered HMM reduce data dimensionality and increase classification performance due to extra flexibility that model time-variant sequences [15][10]. Although, an issue with HMM is EM algorithm with hidden states, the large number of parameters and risk of over fitting with limited data [30].

B) The Dynamic Bayesian Network (DB-N)
The DB-N is a graphical method to indicate conditional dependencies between variables, such as Markov chains [296], with a large set of parameters [102]. This has been applied to dyadic conversation analysis [101][102][14] with the concept of characterizing speaker interactions in terms of emotional transitions probabilities.

An application of DB-N was used to model the conditional dependency between two interacting partners’ emotion states (active and valance) from the IEMOCAP corpus. The dyadic interactions were modeled in relation to temporal emotion dynamics and mutual influence. These were important factors in automatically recognizing emotional states between speakers in a dialog [532].

C) Influence Model (IM)
Asvathiratham introduced the IM as a mathematically tractable model of random interactions on networks [25]. The status of each node in a network evolves over time, based on the current node status and neighboring nodes statuses, according to first order Markov chains. Theoretically the complex phenomena of interactions between chains can be simulated through this simplified model [26].
The IM has been applied to social analysis such as social network interactions [81], group speaker identification [80], debate influences [84] and dialogue classification [83][85]. A study successfully investigated adult-child dialogue classification from turn-taking parameters integrating the IM and GMM [85].

Often applications assume only one person speaks at a time to reduce modeling complexity [81][80], but is not always justifiable. Simultaneous speech is common and can show important differences in parent-adolescent speech styles [82].

Studies reported issues with speakers tending to remain talking or silent for many consecutive states and led to heavily biased IC, converging to a diagonal array (high self influence and zero cross-influence) [85][84][83].

To mitigate the undesirable effect the number of states had been increased with short and long sub-classes of speech and silence [83][85]. Another solution was to non-uniformly resample sequences, replacing long consecutive states by a single meta-step [84]. These strategies improved model sensitivity at the cost of disturbing time-synchronization and removing information on the relative timing of states.

Pan et al. proposed additional dynamic parameters that observe influence change in time instead of restricted to a single time delay [86]. The model was consistent with social influences measures yet was more computationally extensive.

D) Compare Models

The HMM and variations perform extremely well are complex with hidden states, the number of parameters and risk of over fitting with limited data. Hidden states complicate the inference required in the EM algorithm E-step [30].

The main disadvantage of the DB-N, as with the HMM, is a relatively large number of model coefficients. It is difficult to estimate accurately with minimal data and poses the issue of meaningful explanations of the desired solution.
The IM is a simplified HMM, with less parameters and no hidden states, with lower modeling power. IM applications have been limited to first order turn-taking analysis based on speech and silent states. However, due to a large level of ambiguity, it has limited behavioral modeling capabilities [84][86][83][85][80][81].

Experiments with natural speech pointed that 1st order Markov models are not capable of handling long-time steady states of turn taking [86]. Longer delays between subsequent states are more important in emotional and psychological analysis [87][88]. Introducing higher order complex states to describe behavior could provide meaningful psychological interpretations and emotional influence patterns.

2.4.4 Summary Conversation Analysis Depression Recognition Literature
Depression has been examined by multiple research fields with many approaches, and the problem in still ongoing. With an immense number of restricting factors combined efforts of multiple fields (engineering and psychology) are needed to develop efficient diagnostic, analysis and therapeutic strategies.

There is a notion that interactions between speakers exhibit characteristics related to behavioral properties especially in parent-adolescent conversations. Mental health issues have been studies by psychologist and social science and found problems, including depression, can be correlated to speech commination.

In this application the idea is to enhance the IM to include additional emotional construct states introducing an Emotional Influence Model (EIM). Furthermore, the standard first order EIM is extended to dynamic (DEIM) and higher order (HOEI) models. The EIM parameters can be used for qualitative risk assessment and in quantitative machine learning algorithms to detect depression. To our knowledge no studies have focused on a CMS approach for automatic depression detection.
2.5 Literature for Emotion Time Series Forecasting

The aim of this section is to provide the motivation for emotion prediction/forecasting of dyadic conversations with a review of possible models capable of forecasting and a brief overview of related studies including affect studies. Emotion prediction is suitable to have information related to affect states of speakers ahead of time, which could be useful for analysis of conversations of depressed participants.

2.5.1 Methodologies and studies related to time forecasting

The aim of time series modeling is to use past observations to develop a model for prediction/forecasting. This is important in many fields such as economics [239][148][240][241], weather [242] and speech recognition [212][217], social analysis/networks [243][244] and emotional forecasting [248].

Many methods exist for time series modeling for forecasting, starting with development in the 1950s of exponential smoothing models [277][272], exponential weighted smoothing [273], exponential weighted moving average [274] and damped exponential smoothing [275]).

In the 1960s Kalman proposed processes using Kalman filtering [281]. This has been a popular method that describes Markov process with a linear quadratic estimator towards time series forecasting[280].

A popular time series model is the Autoregressive integrated moving averages (ARIMA) [266][270][271][277] and subcases including: autoregressive (AR) [267][277], moving average (MA) [267][277] and moving autoregressive moving average (ARMA) [267][277] describe data based on autocorrelations and moving averages. The theory of ARIMA led to the systematic procedure for application using Box–Jenkins models [267][268] but has issues with non-stationary sequences [269].
An issue is the requirement of a large amount of historical data in order to produce accurate results. Another limitation is the models are inherently linear and non-linear characteristics are missed. These methods had been replaced with the rise of computer power and more sophisticated forecasting methods.

Representing future time series values as a linear combination of past values is easy to implement and may provide good approximations in some relatively simple applications. However it is not adequate or powerful enough in nonlinear real-life systems and performs poorly with complex systems [41]. Often, a non-linear autoregressive model (NAR) is more effective compared to linear AR [252][253].

NN have been used to generate NAR (NAR-NN) in many applications [255][241][240] broadly reviewed in [76]. NAR-NN is computationally powerful and learning is more effective than other NN [115][77][78]. NAR-NN with external input (NARX-NN) has embedded memory, reduces sensitivity to long-term dependencies, increases learning capability and generalization [79][115] and outperforms NN [250][251][249]. Recurrent NN are considerably slower than feed forward NARX-NN and pattern recognition is superior to function approximation [148].

Neuro-fuzzy systems have recently become popular in predicting non-linear time series[39][41][42][43][44]. NN have difficulty interoperating how a decision was made, but due to back propagation is adaptable. Fuzzy systems are smoother and have better interoperation but adaptability is low[154][152][40].

ANFIS integrates advantages of NN and fuzzy system resulting in good learning and interoperation capabilities. ANFIS can learn from a smaller training set and converges faster than a NN. In certain cases fuzzy based systems (fuzzy, fuzzy NARX, ANFIS) have outperformed NN (NN, NARX) [239].
Random walks (RW) can explain observed behaviors of stochastic activity and model time series for prediction [51]. The concept has been applied in biology[48], physics[49] [50], chemistry[47] and computer science [46].

Real life social networks can be modeled by RW such as sentiment analysis [52]. A study investigated and positive and negative sentiment classification with 2D RW and reached 93% outperforming baseline methods [45]. Additionally, positive, negative and neutral emotions reached 77.9%.

2.5.2 Time series forecasting relation to affective forecasting
Emotions are important factors affecting dynamics of social interactions and therefore can be used to predict a person’s mood based on historic series of emotional states. When two people are engaged in conversation, their emotional states observed by an independent external assessor can change within seconds.

In psychology, affective forecasting stems from fundamental works by psychologists Wilson et al. [245]. Applications related to these works developed qualitative affect prediction based on self-reported descriptions of emotion states. These approaches were shown to be subjectively overestimated [245][246][247].

A study on predicting emotional responses of online communication used a data set of real-values on two affective dimensions (valence and arousal). Several prediction methods were compared with the Geometric Means approach the best, attaining a Pearson’s r-value of 0.89 in valence and 0.42 in arousal prediction [243].

In [248], a weighted linear sum of historical data was used to predict emotions of four classes of facial expressions: neutral, joy sad, and surprise. The system was optimized using a genetic algorithm and achieved 70% average prediction accuracy.
In [244], a dynamic continuous graph was applied to predict emotions based on past events in real and virtual web-based social network. The technique applied a discriminative model to determine the emotion at time instance based on a single past state at t-1. Experimental results led to 62% average prediction accuracy.

2.5.3 Summary of Emotion Forecasting Literature

Overall there has been minimal research in the application of prediction models towards affect prediction using annotated sequences. There is no research in relation to prediction of affect in dyadic conversations and should be extended for this purpose. Particularly, there no work conducted in prediction of affect from parent-adolescent conversations including depressed subjects. Prediction of interactional tasks could be useful to analyze the characteristics related to behavioral properties exhibited by depressed adolescents.

In past research NAR(X) models showed promising results for the intended applications and will be examined in this thesis for both NN and ANFIS methods for affect prediction of adolescents. The RW is also useful for prediction efficiency and can be extended in this application with an extension to interacting chains of speakers including influence based on the conversation modeling system (CMS) parameters. The RW is also extended to give a visual representation of conversations towards analysis of depression based on the walk drift.
2.6 Summary

Overall there has been extensive research in both emotion and depression fields but is yet to be linked together in an engineering perspective. There is a psychological and physical correlation between depression and emotion that give promise that there is supplementary information in both concepts that could be mutually beneficial.

Fundamental limitations of acoustic analysis, due to recording miss-matches and quality, can be overcome by analyzing depression without recorded speech signals. There is a notion that emotional interactions between speakers exhibit characteristics related to behavioral properties. This concept is used in this thesis to introduce a Conversation Modeling System (CMS) as an extension of the Influence Model (DEIM and HOEIM) to analyze a subject’s depressive status based on characteristics and correlations of emotional sequences.

In this study the sequences are manual annotations to improve the overall CMS performances. Although, it is possible to integrate the CMS with an automatic affect annotations using an AER or affect prediction component. In this work the affect prediction is examined as an extension on NAR(X) and random walks as a visual representation of conversations towards analysis of depression and improved using CMS parameters. The purpose of the AER is to investigate DNN using new 2D images, representing acoustic feature contours, and the extension to an OMC-DNN.

The motivation of acoustic depression detection was based on the idea that depression is correlated to emotions and could use the same notion of acoustic features as a classification method. A new idea is also investigated with the idea that adolescent depression can be detected using parent’s speech, based on the association of conversational interactions and depression. Acoustic depression detection is used as a standard benchmark for a new depression detection system.
Chapter Three:

DATABASE

3.1 Preview

The audio-visual database used for the depression experiments was obtained through collaboration with the Oregon Research Institute (ORI), USA, containing recordings of naturalistic parent-adolescent conversations. This chapter outlines: (1) participants, collection and depression diagnosis, (2) the recording process, (3) conversational setup of the interactive tasks, (4) details of the database’s LIFE annotation and (5) the formulation of the corpus for the depressed and control group in these experiments.

Furthermore, this thesis investigates automatic emotion recognition as a front end to an overall depression related system. In order to compare results of this thesis to past benchmark studies a popular and publicly available database was chosen: The Berlin emotional speech database (EMO-DB) [174]. This chapter provides details on: (1) participants, (2) recoding setup, (3) the emotional sentence structure and (4) description on perceptions tests used to generate the final emotional corpus.

3.2 Depression Database

A fundamental challenge in depression analysis is the acquirement of a suitable dataset. It is difficult to collect a reasonable amount of data with matched clinical assessments. The earliest depression studies were limited to reasonably small databases: 13 participants [104][397], 16 subjects [100][98] or 17 subjects [156].
Some datasets have been larger and importantly contained control subjects: 30 subjects (only 10 depressed) [106], 32 subjects during and after depression [390].

The largest and most clinically validated datasets are collected through collaboration with psychology departments, research institutes and hospitals. This includes datasets collected by the Psychiatry and Behavioral Health department at the Medical College of Georgia (33 subjects) [139][105][478], the Vanderbilt University Emergency Room and Psychiatric Treatment Unit (17 patients) [426], the U.S Department of Veterans Affairs’ online Crisis Hotline (427 recordings) [425] and an ongoing collection by the Black Dog Institute (80 subjects) [421][434][435][422].

Specifically for adolescents minimal databases are available with just two to our knowledge. The ORYGEN-Youth Health Research Center database has 191 adolescent subjects, initially with no mental health problems, with only 15 developing depression after 2 years [65][87][92][454]. The Oregon Research Institute (ORI), 152 participants including 75 depressed adolescents, was collected through collaboration with psychologists and includes synchronized annotations [24].

The database is a major component for this research and needs to meet requirements to be suitable for the development of an adolescent depression detection system for acoustic and conversation based approaches. Obvious restrictions include a limited age range, which removes many of the available databases.

The ORI-DB met requirements of this research as a large annotated database with psychologically validated diagnosis of adolescents with unscripted conversations [96][97]. This is beneficial as spontaneous speech delivers higher depression detection rates than read [422][457][427][435]. ORI annotation [87][320][88][24], video[456] and audio components [138][107][108][140][147] have been validated in psychology and engineering studies; providing a comparison to this thesis.
3.3 Oregon Research Institute (ORI) Database

3.3.1 Participants

The Oregon research institute (ORI-DB) contains 152 adolescents (75 depressed and 77 controls), between 14 and 18 years old, conversing with parents. The process of participant recruitment, assessment procedures, questionnaires, interviews and depression diagnostics are found in [24]. The adolescent is considered depressed if they met the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) criteria for major depressive disorder (MDD)[22] and non-depressed if the diagnostic criteria for any psychiatric disorder was not met and had no history of mental illness.

The participants were obtained from a sample of West Oregon, USA community with the attempt to match the depressed and control groups in a range of demographic variables including age, gender, ethnicity, and the socioeconomic status [24]. Depressed participants generally came from a lower socioeconomic status and had mothers with higher depressive symptom levels; reflecting associations between adolescent depression and socioeconomic status and maternal depression [7].

3.3.2 Recording Setup

The recording setup utilized the following equipment:

- Video equipment: VideoBank™ system
  - Two cameras each directed at a parent or adolescent
  - Directed at the upper body to ensure the face is captured
- Audio equipment: Audio Technica (model: ATW-831-w-a300) lapel wireless microphone
  - Placed at chest level in a fixed position
  - Parent and adolescent recorded on separate stereo channels
  - The sampling frequency used for recording was 44kHz
The recordings were completed in a quiet laboratory room at ORI with participants seated a few feet apart to simulate a normal conversation. The setup was designed to preserve as much of the natural expressed emotions with unscripted conversations [96][97]. Additionally the participants were outfitted with sensors to measure physiological signals including an electrocardiograph, impedance cardiogram, skin conductance, respiratory and blood pressure, and did not impede speech behavior.

### 3.3.3 Conversational data collection procedure

Three distinct types of family interactions, each 20 minutes in length, were designed to access different behavioral characteristics based on the following descriptions:

- **Event planning interaction (EPI):** discusses planning of a vacation with family and then reminisce about a fun experience they shared (positive tasks)

- **Family consensus interaction (FCI):** two discussions related to family life about identifying and describing best and worst years the adolescent had experienced and the most challenging and rewarding parenting aspects (reminiscence tasks).

- **Problem solving interaction (PSI):** During the discussion the family were attempting to resolve the problems, from two previously agreed upon topics of disagreement, and become mutually agreeable over the subject (conflict task).

These tasks were chosen specifically as they have been found to elicit differential levels of happy, angry, and dysphoric affect. More detailed descriptions on the conversational tasks can be found in [96][97].
3.3.4 Database Annotation Living-In-Family Environments Coding System

The Living in family environments coding system (LIFE) annotates conversations and designed to capture affective interpersonal behavior of participants during family interactions (instructions of affect and content coding found in [23][96]).

Trained observers, blind to diagnosis, produced codes with 20% of interactions by different observer pairs and positively assessed for inter-observer agreement [97], with $k=82$ indicating good strength of agreement [256]. LIFE validity has been established in psychology [88][24][87][320] and engineering studies [107].

LIFE is a second-by-second behavioral based coding system that describes 10 emotions (affect codes): contempt, anger, belligerence, neutral, pleasant, happy, caring, anxious, dysphoric and whine and 27 content codes. The annotations are based on visual and audio cues covering both verbal and non-verbal social signals.

Non-verbal cues are important in determining speaker characterization [301][94] based on facial expression, voice, body posture, gaze and movement [96]. This covers how emotions are expressed as suggested by studies [94][95].

For certain emotions audio is dominating in determining the state, other emotions are more visually dominant. Both carry overlapping and unique information about affective state and are complementary not redundant [446]. An annotation system that fuses both sources could possibly improve depression detection [357].
3.3.5 *Speech Corpus – Experimental Group (Depressed and Control)*

The depression detection experiments, Chapters Five and Seven, and emotion prediction experiments, Chapter Six, the ORI-DB was restricted to a sub-set of dyadic conversations with an adolescent and one parent.

The remaining corpus totals 63 subjects including 29 depressed and 34 control adolescents for a total of approximately 63 hours of data (63 sessions * 3 interactions * 20 minutes).

The distribution of genders pairs is given by Table 3-1, and shows a bias towards mothers-daughter. The gender ratio mirrors the recognized trend of higher depression rates in females [307] and a problem with studies [417][138][107][108][140][105][65][426].

<table>
<thead>
<tr>
<th></th>
<th>Depressed</th>
<th>Non-depressed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Daughter</td>
<td>Son</td>
</tr>
<tr>
<td><strong>Mother</strong></td>
<td>21</td>
<td>5</td>
</tr>
<tr>
<td><strong>Father</strong></td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td><strong>29</strong></td>
<td><strong>34</strong></td>
</tr>
</tbody>
</table>

The data used in these studies include the speech signals for acoustic depression detection in Chapter Seven and the LIFE annotation within Influence Models, in Chapter Five towards depression detection. In addition the LIFE annotations were used in investigations related to emotional forecasting/prediction of time series Chapter Six.
3.4 Emotion Database

A challenge of affective computing is to generate or acquire an adequate representation of emotions to develop models appropriate for the application. The main hurdle is the insufficiency of labeled emotion data inhibiting capabilities and performance in practical situations [207][405][406].

Primarily, the nature of emotional databases acted, induced (stimuli) or spontaneous (real environment) is important. Spontaneous speech is more natural and closer to real world application requirements. Acted speech is normally expressed with exaggeration and interoperaed as more intense and prototypical. Acquiring spontaneous speech can be difficult and limited databases are available. Acted data is more readily available as most of the well-annotated databases contain acted corpora.

Another consideration is labeling, as there is no consistent emotion definition and no unified standard of emotion categories. A review by Ververidis et al. found 29 different emotions used in 64 distinctive datasets with most studies limited to between four and six emotions [172]. Ekman proposed 15 basic emotions but admitted this is debatable and only neutral, joy and anger were the most predominant [368]. Cowie et al. advised the “big-six” as key emotions: anger, happy, sad, fear, surprise, and disgust [369][233]. These can be considered archetypal, representing popular emotions, ideal for testing AER capabilities [369] [356].

Acted emotions are the most widely available and improve performance by providing less subjective ground-truths. SEMAINE and IEMOCAP databases only use primitive emotions (e.g. valance and activation) limited to three to five states. EMO-DB and eNTERFACE are advantageous with discrete states (i.e. anger, happiness) of seven or six primary emotions.
3.5 Berlin Emotional Speech Database (EMO-DB)

In this research a popular, publically accessible database with benchmark results was ideal. The EMO-DB has been chosen as one of the most often used databases and thus providing a wide range of comparison with other studies.

3.5.1 Participants

A pre-selection group of 40 German speaking professional actors and performed an audition judged by three experts. The perceived naturalness and recognizability was used to find 10 participants (5 male and 5 female) to record the entire database.

3.5.2 Recording Setup

The recordings were collected in an anechoic chamber, designed to absorb reflections of sound and insulated from exterior noise, to achieve a high audio quality. The participants stood 30cm from a Sennheiser MKH 40 P 48 microphone and recorded with Tascam DA-P1 portable DAT recorder at 48 kHz sampling frequency [174].

3.5.3 Emotional utterance collection procedure

Each participant acted 10 German sentences, with linguistically undefined emotion content with one phrase for five sentences and five with two phrases. 7 emotions were simulated including anger, boredom, disgust, happiness, fear, sadness and neutral speech, some combinations of emotions and sentences repeated, totaling 816 phrases.

3.5.4 Speech Corpus – Experimental Group (Emotions)

20 listeners performed perception tests with samples retained if at least 80% clearly assigned an emotion and more than 60% perceived natural speech. After perception tests 488 phrases were available, in Table 3-2, with 84% listener accuracy [174].

<table>
<thead>
<tr>
<th></th>
<th>Anger</th>
<th>Bored</th>
<th>Disgust</th>
<th>Fear</th>
<th>Happy</th>
<th>Neutral</th>
<th>Sad</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>60</td>
<td>34</td>
<td>8</td>
<td>26</td>
<td>21</td>
<td>38</td>
<td>17</td>
</tr>
<tr>
<td>Female</td>
<td>67</td>
<td>45</td>
<td>30</td>
<td>29</td>
<td>37</td>
<td>40</td>
<td>36</td>
</tr>
</tbody>
</table>

Table 3-2 DISTRIBUTION OF EMOTION IN THE BERLIN EMOTIONAL SPEECH DATABASE (EMO-DB) FOR BOTH MALE AND FEMALE SUBJECTS
Chapter Four:

EMOTION RECOGNITION

4.1 Preview

This chapter explores automatic emotion recognition (AER) with the major components outlined by Figure 4-1. The EMO-DB was pre-processed resulting in voiced frames and acoustic speech features extracted including MFCC, TEO and derivatives, comparing features derived from glottal and speech waveforms.

The DNN was trained with the proposed 2D images representing high-level features depicting contours of the feature vectors. Multiple classifiers including Support Vector Machines, Neural Networks and Gaussian Mixture Models were examined as benchmarks to determine if the new DNN method was beneficial and if any other classifier could efficiently use the 2D features compared to feature vectors.

An Optimized Multi-Channel Deep Neural Network (OMC-DNN) was proposed towards AER. A novel aspect of the methodology is a parallel configuration of single DNN combined with weighted late integration at the decision level. The weights are optimized using a hybrid Simulated Annealing Genetic Algorithm optimizer.
The methodology is explained in Sections 4.3, 4.4, 4.5, 4.6 and 4.7 for preprocessing, feature extraction, image generation, classifiers and the OMC-DNN. Section 4.8 and Section 4.9 outlines the classifiers and experimental setups and Section 4.10 provides the evaluation methods. The results are supplied in Sections 4.11 and 4.12 and a discussion and summary in Section 4.13.

### 4.2 Automatic Emotion Recognition

The general framework of the AER system for training and testing stages are given in Figure 4-2 and Figure 4-3. Both stages involved data preprocessing and the resultant windowed frames then used to extract acoustic features. In the training stage the features were used to learn models for each emotion class. The testing stage used the learned models to classify the emotion of a test. The main steps: preprocessing, feature extraction and classification are explained in the subsequent subsections.

![Figure 4-2 General overview of emotion recognition training stage for N emotion models (binary or multiclass) including a training set, pre-processing, feature set extraction and classification system](image1)

![Figure 4-3 General overview of emotion recognition test stage for N emotion models (binary or multiclass) including a test sample, pre-processing, feature set extraction and classification using the trained emotional models](image2)
4.3 Pre-Processing

The speech from the EMO-DB was pre-processed starting with amplitude normalization of the signals based on the maximum absolute amplitude given by equation (4.1), where $x(n)$ donates the amplitude of the speech sample.

$$x_N(n) = \frac{x(n)}{\max |x|} \quad \text{(4.1)}$$

Speech was segmented into 16ms windowed frames with a 50% overlap using a Hamming window to increase time resolution and avoid discontinuities between segments. The Hamming filter is a preferable window as it minimizes the magnitude of the nearest side-lobe in the frequency domain. If the final frame was too short it was appended with random noise, 30dB below the maximum amplitude of the frame.

4.3.1 Voiced Activity Detector (VAD)

Speech production is categorized as voiced or unvoiced and only voiced of interest in speech analysis purposes [326][327]. This is understood based on physiological production of speech, as explained in Section 2.2.2.

Voiced speech is produced by airflow through the glottis and vocal cord vibrations exciting the vocal tract. Unvoiced speech is generated by air forced through a constricted vocal tract and does not vibrate vocal cords and hence has no fundamental frequency.

The goal was to extract voiced speech segments and remove unwanted unvoiced and silent segments using a VAD. This was achieved with linear prediction [326][327] implemented from the MATLAB speech processing and synthesis toolbox [184].
Initially, speech was filtered to remove low frequency drift using a zero-phase forward and reverse filter by a process of filtering, reversing, filtering and reversing [324]. The filter output, $Y(m)$, is given by equation (4.2) where $X(m)$ is the input and $A$ and $B$ are the recursive and non-recursive coefficients.

$$
Y(m) = A(1) \times X(m) + B(2) \times X(m - 1) + \ldots + B(mb + 1) \times X(m - mb) - A(2) \times Y(m - 1) - \ldots - A(mA + 1) \times Y(m - mA)
$$ (4.2)

After filtering, the 13th order linear prediction coefficients, the energy of the prediction error and the first reflection coefficient were calculated. The first reflection coefficient, $r_1$, is given by equation (4.3) where $x(n)$ is the $n^{th}$ for $N$ frames.

$$
r_1 = \frac{1}{N} \sum_{n=1}^{N-1} x(n)x(n + 1) - \frac{1}{N} \sum_{n=1}^{N} x(n)x(n)
$$ (4.3)

The energy, $E$, of a frame was calculated as the geometric mean of the energy of two sections of a frame, given by equation (4.4), where $x_1(n)$ is the first half of the frame and $x_2(n)$ is the second half of the frame. This is a conventional definition modified to ensure the values independent on the signal dynamic range.

$$
E = \sqrt{x_1(n)^2 \times x_2(n)^2}
$$ (4.4)

The threshold of $r_1$ and $E$ was experimentally and empirically determined by previous studies [147][169]. If $r_1>0.2$ and $E>1.85*10^7$ the frame was defined as voiced otherwise it was unvoiced or silent and removed, shown by Figure 4-4.
Figure 4-4 Illustration of the voice activity detector from left to right, top to bottom (a-d) correspond to: (a) Entire speech signal (b) Concatenation of the extracted voiced frames (c) and (d) spectral-domain for original signal and after VAD

### 4.3.2 Glottal Waveform Estimation

During speech production the airflow through the glottis causes vibrations of the vocal folds, generating the periodic glottal waveform, which is then modulated (filtered) by the vocal tract. Remembering, the source-filter model in Section 2.2.2, speech production is represented as a linear-model of an excitation source $E(z)$, with glottal, $G(z)$, vocal tract, $V(z)$, and lip radiation filters, $R(z)$, as follows:

$$y(n) = e(n) * g(n) * v(n) * r(n)$$  \hspace{1cm} (4.5)

$$Y(z) = E(z)G(z)V(z)R(z)$$  \hspace{1cm} (4.6)
Speech is the convolution of the glottal waveform and impulse response of the vocal tract. The glottal waveform, \( u_g(n) = e(n) * g(n) \), was determined by inverting the estimated vocal tract transfer function \([66][67]\), given by equation (4.7). Inverse filtering is a popular method to cancel out the effect of vocal tract filtering \([305]\).

\[
U_g(z) = \frac{Y(z)}{V(z)R(z)} \tag{4.7}
\]

In this application, the glottal waveform was extracted using the TKKAparat glottal inverse filtering toolbox \([185]\). The glottal inverse filtering was established with an iterative adaptive inverse filtering algorithm (IAIF) \([66]\) and a discrete all-pole model (DAP) \([66][67]\). The DAP has less sensitivity to biasing of formants, produced from nearby harmonic peaks, compared to LP filter when modeling the vocal tract \([140]\). The following steps summarize the IAIF process:

- Estimate the vocal tract filter using 1st-order discrete all-pole model (DAP)
- Remove the effect of the vocal tract with inverse of vocal tract filter
- Result is time-domain waveform of estimated glottal source
- Process repeated with a higher order DAP to improve glottal estimate

The IAIF decomposes speech into the vocal tract and glottal transfer function assuming the glottal excitation contribution is estimated by a low-order DAP \([66][67]\). Each frame, \( y(n) \), was filtered with a low order DAP to estimate the vocal tract, \( V(z) \), defined by equation (4.8) where \( a_k \) are filter coefficients and \( p \) the number of poles. The glottal transfer function, \( U_g(z) \), is the inverse of \( V(z) \) with a gain factor of \( G \) given by equation (4.9) and repeated for higher order DAP iteratively.

\[
V(z) = \sum_{k=0}^{p} a_k Z^{-k}, \quad a_0 = 1 \tag{4.8}
\]

\[
U_g(z) = \frac{G}{V(z)} \tag{4.9}
\]
4.4 Feature Extraction

Acoustic features include MFCC and TEO generated from the speech (S-MFCC and S-TEO) and glottal (G-MFCC and G-TEO) waveforms and corresponding first (ΔS-MFCC, ΔG-MFCC, ΔS-TEO, ΔG-TEO) and second (ΔΔS-MFCC, ΔΔG-MFCC, ΔΔS-TEO, ΔΔG-TEO) derivatives. The features are outlined in the following subsections.

MFCC and TEO have been robust in AER applications and show promising results when derived from the glottal waveform (augmented) [238][65][147][342][367], improvement with derivatives (modified) [138][518] and as a vector could be reformed as a 2D image (representation).

4.4.1 Cepstral-Mel frequency cepstral coefficients (MFCC)

In the cepstral domain the vocal cord excitation and vocal tract effects are separable which is effective for speech analysis [350][351]. Humans perceive small changes in pitch easier to discern at low frequencies than at high frequencies. Cepstral analysis and features has been extended to reflect the perceived non-linear hearing response.

![Figure 4-5](image)

Figure 4-5 a) Mel-Scale versus frequency (Hz) scale conversion and b) triangular Mel filter-banks with non-linear separation on the frequency scale

Mermelstein is credited with creating the Mel-scale, which is experimentally derived based on perceptual pitches, that simulates perceived pitch to reflect hearing response [56][57][58].
Conversion between the linear frequency scale, $F_{\text{linear}}$, and the Mel-scale, $F_{\text{mel}}$, is shown by Figure 4-5 and defined by equation (4.10). The corner frequency, $C$, is 700 as it provides the closest approximation to the Mel-scale [59].

$$F_{\text{mel}} = \frac{C \log_{10} \left( \frac{1 + F_{\text{linear}}}{C} \right)}{\log_{10} 2}$$  \hfill (4.10)

The Mel frequency Cepstrum represents the short-term power spectrum based on the cosine transform of the log power spectrum on the Mel scale [56][57]. For each frame the Mel frequency cepstral coefficients (MFCC) are calculated as follows:

- Calculate the Fourier Transform of a windowed frame
- Map the power spectra, $S$, to the Mel-scale using triangular filter bank, $H_i$,
- Take the logarithm of the powers at each Mel frequency, $Y_i$
- Sum the energy in each filter
- Compute the discrete cosine transformation of the log filter bank energies
- Amplitude of the DCT coefficients result in the MFCC

Overlapping triangular filters were placed linearly on the Mel-scale ($F_{\text{mel}}$) and then mapped back to the linear frequency scale ($F_{\text{linear}}$), shown in Figure 4-5b. These Mel-frequency filter banks were applied to the power spectrum, $S$, equation (4.11), where $H_i$ is the $i^{th}$ triangular filter and $m_i$ and $n_i$ give the filter boundaries.

$$Y(i) = \sum_{j=m_i}^{n_i} \log_{10}[S(j)] H_i(j)$$  \hfill (4.11)

The $n^{th}$ Mel-frequency cepstral coefficient was calculated by taking the DCT defined by equation (4.12). Where the center frequency of the $i^{th}$ filter is given by $k_i$, $N$ is the number of filters and $NFFT$ is the number of samples in the FFT.

$$MFCC(n) = \frac{2}{N} \sum_{i=1}^{N} Y(i) \cos \left( k_i \frac{2\pi}{NFFT} n \right)$$  \hfill (4.12)
4.4.2 Teager energy operators (TEO)

Spectral, cepstral, prosodic and glottal features assume speech production is a linear model. In the 1980s, Teager et al. argued a linear speech model was an inaccurate simplification for a speech production model and required non-linear modeling [69].

Teager et al. further explored non-linear speech characteristics and discovered evidence during each glottal cycle the non-linear air flow causes turbulence known as vortices [68]. Many other experiential studies have established strong evidence that glottal airflow is non-linear, through vocal folds with laminar components leading to vortices [70]. Vortices provide additional excitation signals flows that produce additional harmonics and cross-harmonics in the speech spectrum [55][464][70].

Teager proposed parameters that model the time-varying vortex flow as a non-linear energy-operator called Teager energy operators (TEO), $\Psi$ [68][69][70]. For a continuous signal, $x(t)$, TEO, $\Psi$, is given by equation (4.13), where $\dot{x}(t)$ and $\ddot{x}(t)$ are the signal first and second derivatives. Kaiser proposed a discrete form, defined by equation (4.14), as an estimate of instantaneous energy in $x[k]$/[53][54].

\[ \Psi(x(t)) = \dot{x}^2(t) - x(t)\ddot{x}^2(t) \]  
(4.13)

\[ \Psi(x[k]) = x^2[k] - x[k + 1]x[k - 1] \]  
(4.14)

Assuming speech is amplitude and frequency modulated signal (AM-FM), TEO parameters represent the instantaneous energy of a signal as a function of both amplitude and frequency resembling an energy-profile [345][55]. This indicates TEO characterize spectral energy distribution and have high sensitivity to transient changes that assist in detecting additional harmonics and cross-harmonics [140][55][70].
4.4.2.1 Area under TEO critical band autocorrelation envelopes (TEO-CB-Auto-Env)

The auditory system includes a filtering process and the audible frequency range is segmented into frequency bands [361], equivalent to a bank of band-pass filters [362]. Zhou et al. proposed calculating TEO in frequency bands to help observe additional harmonics [55] and has been used successfully in studies [238][147][345].

Figure 4-6 Wavelet Packet Tree (WPT) with three levels of decomposition using low pass filter denoted by \( L \) and the high pass filter by \( H \). The filter coefficients for approximation and detail are the approximation and detail coefficients.

In this application the signal was decomposed into sub-bands using wavelet packets, a modified DWT with twice the number of filters [304]. Wavelet packet trees, WPT, iteratively filter the speech with both high and low pass filters, shown by Figure 4-6. At each decomposition level the approximation and detail coefficients from the previous level are filtered. The WPT contains a set of Wavelet Packet Coefficients, WPC, that provide 2D frequency-time representations for multiple frequency bands. 17 frequency bands were extracted, denoted by Table 4-1, using WPC to a depth level of 5 using Coiflet wavelet packets [302] and Shannon entropy.

<table>
<thead>
<tr>
<th>Table 4-1 TEO SUB-BAND’S UPPER (U), LOWER (L) FREQUENCIES (HZ) AND BANDWIDTH (BW) DERIVED FROM THE WAVELET PACKETS</th>
</tr>
</thead>
<tbody>
<tr>
<td>17 TEO Critical Bands Determined with Wavelet Packets</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>-------------------------</td>
</tr>
<tr>
<td>L</td>
</tr>
<tr>
<td>U</td>
</tr>
<tr>
<td>BW</td>
</tr>
</tbody>
</table>
Zhou et al. proposed a TEO variation by calculating the area under the normalized TEO autocorrelation envelope, shown by Figure 4-7, for each critical band (TEO-CB-Auto-Env) [55]. Speech analysis is more robust using characteristics that are derived from the area under the normalized TEO autocorrelation [55]. The area under the normalized TEO autocorrelation envelope is defined by equation (4.15), where $M$ is the number of samples and $L$ is the correlation lag.

$$R_{\psi(L)} = \frac{1}{2M + 1} \sum_{n=-M}^{M} \psi(x[n])\psi(x[n+L])$$ (4.15)

Given a signal with a single harmonic and constant instantaneous amplitude and frequency the area under the normalized TEO autocorrelation function is equal to $M/2$ [238]. Speech frames with multiple harmonic components produce a time-varying autocorrelation envelope that decays to zero. The area under the contour is now less than $M/2$, indicating the TEO-CB-Auto-Env detects changes in harmonics.

Figure 4-7 Flow chart outlining the process of calculating the TEO_CB_AUTO_ENV coefficients from voiced speech frames. The derivation involves generating the TEO for the 17 sub-bands, found from the WPC, and then calculating the area under the autocorrelation contour.
4.4.3 Glottal Waveform Features

Glottal waveform characteristics have widely been studied and improved speech recognition tasks [63][64]. Deriving features from the glottal waveform instead of the speech waveform is valuable to improve recognition tasks [147][238][65][342][367]. In this research the following features were derived from the glottal waveform:

1. Mel-frequency Cepstral Coefficients from the glottal waveform (G-MFCC)
2. Teager Energy Operator from the glottal waveform (G-TEO)

The MFCC and TEOs, derived from the glottal waveform, follow the same methodology used from the speech waveform, as in Sections 4.4.1 and 4.4.2.

4.4.4 Delta (Δ) and Delta-Delta (ΔΔ) Coefficients

Dynamics features such as first- and second-order derivatives (delta and delta-delta) capture temporal information from neighboring frames with difference coefficients. The delta (Δ) and delta-delta (ΔΔ) coefficients are calculated using equation (4.16) and (4.17), for the previously mentioned S-MFCC, G-MFCC, S-TEO and G-TEO.

\[
\Delta_t = \frac{\sum_{\theta=1}^{\phi} \theta (c_{t+\theta} - c_{t-\theta})}{2 \sum_{\theta=1}^{\phi} \theta^2}
\]  

(4.16)

Where \( \Delta_t \) is the delta coefficient at time \( t \) of the corresponding static coefficients from \( c_{t,\theta} \) to \( c_{t+\theta} \), the window size is defined as \( \Theta=9 \). Similarly, applied to the derivatives, \( \Delta_{t,\theta} \) to \( \Delta_{t+\theta} \), to obtain the second derivative, \( \Delta\Delta_t \).

\[
\Delta\Delta_t = \frac{\sum_{\theta=1}^{\phi} \theta (\Delta_{t+\theta} - \Delta_{t-\theta})}{2 \sum_{\theta=1}^{\phi} \theta^2}
\]  

(4.17)
4.5 2D image generation

Existing emotion classification techniques ordinarily represent input parameters (features) as acoustic speech feature vectors. This research proposed 2D secondary features representing vectors as image contours, shown by the Figure 4-8 examples.

Figure 4-8 Examples of 2D (34x34) image representations of acoustic feature sets as a contour (white) given for 12 MFCC coefficients (left) and 17 TEO coefficients (right). Where the x-axis represents the coefficient index and the y-axis represents the normalized range of coefficient values.

The 34x34 image is a matrix created by interpolating 12 MFCC or 17 TEO parameters with 34 quantized points. The interpolated contour is mapped to the matrix with white pixels representing a value of 0 and black pixels value of 1. The x-axis is aligned with the coefficient index and the value of the contour is mapped into 34 quantized levels to the normalized y-axis. The neurons in the input layer encode the 2D images at a pixel basis; such that the input layer has 34*34=1156 neurons (pixels).

4.6 Modeling and Classification systems

The acoustic features extracted from the speech signal are useful for statistical modeling and machine learning for discriminating and classifying classes. This research concentrated on emotion recognition using classification of acoustic features.

The following subsections give an overview of previously established, classification systems, which are used for the benchmark: Support Vector Machine (SVM), Gaussian Mixture Model (GMM) and Neural Network (NN). This is followed by a detailed examination on DNN and concludes with the proposed OMC-DNN.
4.6.1 Gaussian Mixture Model (GMM)

A Gaussian Mixture model (GMM) is a generative classifier widely recognized for effective speech and emotion classification [364][207]. The Gaussian distribution for a single dimension variable, \( x \), is given by equation (4.18), \( N(x \mid m, \sigma^2) \), is a normal distribution defined as the probability of \( x \) given \( \mu \) mean and \( \sigma^2 \) variance.

\[
N(x \mid \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left[ -\frac{1}{2\sigma^2} (x - \mu)^2 \right] \tag{4.18}
\]

Considering \( x \) as a D-dimensional vector an extension to multivariate Gaussian distribution is defined by equation (4.19). Now, \( \mu \) represents a D-component mean vector and \( \Sigma \) is a \( D \times D \) covariance matrix. Instead of a full covariance matrix a diagonal covariance matrix is used for computational efficiency.

\[
N(x \mid \mu, \Sigma) = \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \exp \left[ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right] \tag{4.19}
\]

To model data a weighted linear combination of \( M \) GMM clusters is implemented given by equation (4.20). This describes the probability density function of data \( x \) as a mixture of weighted Gaussians, where \( w_m \) are the mixing coefficients.

\[
p(x) = \sum_{m=1}^{M} w_m N(x \mid \mu_m, \Sigma_m) \tag{4.20}
\]

Each Gaussian density has a corresponding mixing coefficient, mean \( \mu_m \) and covariance \( \Sigma_m \). The individual Gaussian mixture components are normalized and constrained such that \( \sum_{m=1}^{M} w_m = 1 \) where \( 0 \leq w_m \leq 1 \). The aim is to optimize the parameters, \( w \equiv \{w_1, \ldots, w_M\}, \mu \equiv \{\mu_1, \ldots, \mu_M\}, \Sigma \equiv \{\Sigma_1, \ldots, \Sigma_M\}, \) by maximizing the likelihood function given in equation (4.21), for a given \( X=\{x_1, \ldots, x_N\} \) dataset.
\[
\ln p(X|w, \mu, \Sigma) = \sum_{n=1}^{N} \ln \sum_{m=1}^{M} w_m N(X_n | \mu_m, \Sigma_m)
\]  

(4.21)

This is accomplished using Expectation Maximization (EM) optimization algorithm with of two stages: the expectation (E) step and the maximization (M) step. The E-step computes the expectation of the log-likelihood (LL) from equation (4.21) using the current latent variable estimate. Then the M-step re-estimates parameters by maximizing the expected LL found in the E-step using equation (4.22). Alternating between the E-step and M-step iteratively improves the LL until convergence is met.

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

(4.22)

The HTK package, designed for HMM, generated the equivalent to GMM using a one state self-loop continuous density HMM [433]. The EM algorithm within the HTK package is implemented with a forward-backward algorithm (Baum-Welch).

GMM incorporates Bayesian decision procedure to determine the class of a test sample, determined as the highest likelihood, \(p(c_j|x)\). [470]. The Bayes rule, given by equation (4.23), states the probability test vector \(x\) belongs to class \(c_j\) is defined as \(p(c|x)\). Where \(p(x|c_j)\) denotes the posterior probability, \(p(c_j)\) is the priori probability of class \(c_j\), and \(p(x)\) is the marginal likelihood.

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

(4.23)

A disadvantage of GMM is that it fails in higher dimension problems due to computational expense and there is an unclear choice of mixture sizes.
4.6.2 Support Vector Machines (SVM)

Support vector machines (SVM) are a discriminative classifier used extensively in speech classification [409][207]. SVM are considered a state-of-the-art classifier with good generalization [408]. SVM use supervised learning to separate data amid multidimensional hyperplanes (decision boundaries) with maximum class separation.

In 1963, Vapnik et al. outlined a pattern recognition problem [287] and in 1974 developed further as a statistical learning theory [411]. This led to the SVM algorithm proposed by Vapnik in 1979 [410].

In 1992, Boser, Guton and Vapnik proposed a non-linear SVM with a kernel and efficient learning algorithms [288]. In 1995, Cortes and Vapnik created the modern standard SVM with a soft-margin hyperplane [289]. An overview of SVM follows with more details in [285][288][289][408][409][309].

- **Linear SVM**

  Given a dataset \([x_i,y_i], i=1,2,...,N\), where \(N\) is the number of samples and \(x_n \in \mathbb{R}^d\) is a \(d\)-dimensional feature vector with corresponding binary class, \(y_n \in \{-1;+1\}\), the class label pairs are given as \((x_1,y_1), (x_2,y_2), \ldots,(x_N,y_N)\). The SVM learns the mapping, \(\tilde{x}_n \mapsto y_n\), between the feature vectors, \(\tilde{x}_n\), and the true class label, \(y_n\).

  The SVM categorizes classes by separating data with a linear hyperplane, shown in Figure 4-9, that defines the decision boundary. The aim is to maximize the margin, \((d^+ + d^-)\), between the hyperplanes (decision surfaces), \(H_1\) and \(H_2\).
Figure 4-9 Illustration of linearly separable data to train a SVM with ‘+’ and ‘*’ indicating points for each class. The hyperplane, \( H \), separates the two classes with a maximal margin, \( 2/||w|| \), between the parallel hyperplanes \( H_1 \) and \( H_2 \). The data points that lie on the hyperplanes, denoted as circled points, are the support vectors.

The points on the decision surfaces are the support vectors, for which \( \vec{x} \cdot \vec{w} + b = 1 \) or \( \vec{x} \cdot \vec{w} + b = -1 \), and directly effect the optimal hyperplane defined by:

\[
\vec{x} \cdot \vec{w} + b = 0 \tag{4.24}
\]

The decision function determines the output, \( f \), given by equation (4.25) where \( \vec{w} \in \mathbb{R}^d \) is the normal vector representing weight and \( b \) is a real number known as the bias. The SVM decides the function that best approximates the unknown \( y=f(x) \).

\[
f(\vec{x}, \{w, b\}) = \text{sign}(\vec{x}_n \cdot \vec{w} + b) \tag{4.25}
\]
- **Hard margin**

The goal is to determine the parameters, \( \mathbf{w} \) and \( b \), that maximize the margin, \( \frac{2}{\|\mathbf{w}\|} \), with no interior data points [411]. This is equivalent to minimizing the constrained objective problem in equation (4.26), subject to (4.27).

This is easier since it becomes a quadratic programming problem (optimization of quadratic function of several variables subjected to linear constraints).

\[
\min_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{2} \|\mathbf{w}\|^2
\]

subject to \( y_n (\mathbf{x}_n \cdot \mathbf{w} + b) \geq 1, \quad n = 1, \ldots, N \) (4.27)

- **Soft margin extension**

Data with overlapping distributions among classes are not perfectly linearly separable, so a hard margin results in poor generalization capabilities. A soft margin relaxes constraints letting some data points be misclassified. Equation (4.26) is modified with a slack variable \( \xi_n \geq 0 \) and a penalty parameter (regularization) \( C \geq 0 \) to control misclassification tolerance and contribution of data outside the hyperplane [289].

\[
\min_{\mathbf{w} \in \mathbb{R}^d, \xi_n \in \mathbb{R}} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{n=1}^{N} \xi_n
\]

subject to \( y_n (\mathbf{x}_n \cdot \mathbf{w} + b) \geq 1 - \xi_n, \quad n = 1, \ldots, N \) (4.29)
• **Nonlinear SVM**

For classification problems with complex nonlinearities a linear decision surface does not exist. The data is mapped into a higher dimension feature space, $F$, using a nonlinear map, $\varphi: \mathbb{R}^d \rightarrow F$. This transforms the data, $x \rightarrow \varphi(x)$, into a linearly separable space and can easily be classified with a linear hyperplane. The formulation of the quadratic programming problem replaces $x_n$ with $\varphi(x_n)$, where $\varphi$ specifies the higher-dimensional mapping, giving the standard SVM formulation as follows:

$$
\begin{align*}
\min_{w \in \mathbb{R}^d, \xi \in \mathbb{R}} & \quad \frac{1}{2} \|w\|^2 + C \sum_{n=1}^{N} \xi_n \\
\text{subject to} & \quad y_n (\varphi(x_n) \cdot w + b) \geq 1 - \xi_n, \quad n = 1, ..., N \quad (4.30)
\end{align*}
$$

**Lagrange**

In order to cater for the constraints the optimization problem is reformulated by allocating Lagrange multipliers, $\alpha_n$, represented in either primal or dual form and solved analytically [288][289][408]. The dual representation is given by the quadratic optimization problem in equation (4.32) and constrained by equation (4.33).

$$
\begin{align*}
\max_{\alpha} & \quad \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n,m=1}^{N} \alpha_n \alpha_m y_n y_m \varphi(x_n) \cdot \varphi(x_m) \\
\text{subject to} & \quad \sum_{n=1}^{N} \alpha_n y_n = 0, \quad 0 \leq \alpha_n \leq C, \quad n = 1, ..., N \quad (4.32)
\end{align*}
$$
Kernel Trick

The higher-dimension feature space requires the calculation of $\phi(\tilde{x}_n) \cdot \phi(\tilde{x}_m)$, which may be intractable. The solution is to use a special Kernel function that operates on the lower dimensions, $x_n$ and $x_m$, including the following examples:

- Linear function: $K(\tilde{x}_n, \tilde{x}_m) = \tilde{x}_n \cdot \tilde{x}_m$
- Polynomial function: $K(\tilde{x}_n, \tilde{x}_m) = (r + \gamma \tilde{x}_n \cdot \tilde{x}_m)^d$ where $d>0$, $\gamma > 0$
- Radial basis function: $K(\tilde{x}_n, \tilde{x}_m) = \exp(-\gamma \|\tilde{x}_n - \tilde{x}_m\|^2)$, where $\gamma > 0$
- Neural network: $K(\tilde{x}_n, \tilde{x}_m) = \tanh(p_1 \tilde{x}_n \cdot \tilde{x}_m + p_2)$, where $p_1>0$ and $p_2<0$

A kernel function $K(x_n, x_m)$ and produces an equivalent to the dot product of the higher-dimension vectors, as given by equation (4.34), known as the Kernel trick.

$$K(x_n, x_m) = \langle \phi(\tilde{x}_n) \cdot \phi(\tilde{x}_m) \rangle \quad (4.34)$$

So the optimization problem for a non-linear case using the Kernel trick, is transformed in dual form as follows:

$$\max_{\alpha} \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n,m=1}^{N} \alpha_n \alpha_m y_n y_m K(\tilde{x}_n \cdot \tilde{x}_m) \quad (4.35)$$

subject to $\sum_{n=1}^{N} \alpha_n y_n = 0$, $0 \leq \alpha_n \leq C$, $n = 1, \ldots, N$ \quad (4.36)

The free parameters are bounded by the number of support vectors (not variables), which is computationally tractable in high-dimension feature-spaces. Dual representation only $\alpha$ is learned compared to a $d$-dimensional space in primal form. The primal representation solves $(w_n, n = 1, \ldots, N)$, where $N$ is the number of feature and the dual solves ($\alpha_n$, $n = 1, \ldots, N$), where $N$ is the number of samples.
○ Decision function

The dual representation of the decision function from equation (4.25) is expressed using the Kernel trick [285]. This simplifies the expression to a linear combination of points as a function of just the inner product of two points, given by equation (4.37) where \( \alpha \) is a \( N \)-dimension weight vector and \( \tilde{w} \) and \( b \) are given by equation (4.38).

$$f(\tilde{x}, (\tilde{w}, b)) = \text{sign} \left( \sum_{n=1}^{N} \alpha_n y_n K(\tilde{x}_n, \tilde{x}_m) + b \right)$$  \hspace{1cm} (4.37)

$$\tilde{w} = \sum_{n=1}^{N} \alpha_n y_n \varphi(\tilde{x}_n) \text{ and } b = \varphi(\tilde{x}_n) \cdot \tilde{w} - y_n$$  \hspace{1cm} (4.38)

After the parameters are determined, \( \alpha_n, w, b \), considering \( \alpha_n \) is only non-zero when \( \varphi(\tilde{x}_n) \) is on or near the boundary (support vectors) they alone specify the decision boundary. Therefore after training all other data points can be ignored and only need to sum over the \( \tilde{x}_n \) which represent the support vectors [288][289].

○ Optimization/ computing SVM

The SVM is convex quadratic programming (QP) optimization problem and the parameters can be optimized as described in [279][285][286]. Modern approaches use sub-gradient descent, which is efficient with large training sets, or coordinate descent good with large feature spaces [279][290] or sequential minimization optimization [292]; used in MATLAB and LIBSVM toolboxes [291].
4.6.3 Neural Network

A brief explanation of background, terminology and principles of Neural Networks will help comprehend and appreciate the Deep Neural Network discussed later. McCulloch and Pitts created an analogy between biological neurons and simple logic gates with binary outputs [209]. An output, $z$, is produced based on the weighted sum, $w_1, w_2, ..., w_n$, of the inputs $x_1, x_2, ..., x_n$ as illustrated by Figure 4-10.

\[
z = w_0 x_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n = \sum_{j=1}^{n} x_j w_j
\]

Figure 4-10 Perceptron with $n$ binary inputs, $x_i$ and corresponding weights, $w_i$, used to produce a binary output, $z$, as a linear weighted sum of the input vector.

The decision at the output is binary and is defined by the following unit step activation function $f(z)$, where $b$ are the bias values:

\[
output = f(z) = \begin{cases} 
0, & \text{if } \sum_j w_j x_j + b \leq 0 \\
1, & \text{if } \sum_j w_j x_j + b > 0
\end{cases}
\]

Rosenblatt developed supervised perceptron learning rules to find weights and biases [210]. A small weights change results in a large output change and is difficult to control. Besides discrimination is only ensured if the classes are linearly separable.

Artificial neurons, shown in Figure 4-11, relate the output to a transfer/activation function, $f(s)$, such as a sigmoid, logistic or hyperbolic tangent. The advantage is a small change in parameters produces a small change in the output, $y$. 

89
A network of artificial neurons operating in parallel creates a NN with input, hidden and output layers, shown by Figure 4-12. This is a multilayer perceptron (MLP), even though it consists of neurons not perceptrons. In this application a feed-forward architecture is used so outputs are only sent forward not fed-back.

Figure 4-12 Neural Network architecture showing the connections of the artificial neuron nodes, which individually correspond to Figure 4-11, between the input, hidden and output layers.

The weights and biases of the network are learned so the output is close to the actual values of the training data. To measure how well the network is trained the aim is to minimize an objective function, which can be the mean square error (MSE), achieved through gradient descent or other optimization algorithm.
4.6.4 Deep Neural Network (DNN)

The architecture of a DNN, shown by Figure 4-13, has a NN structure with \( L \) hidden layers \((h)\). The visible input layer, \((v)\), is related to the structure of the features, and the output layer dependent on class size. The weights, \( W^{(j)} \), between the units of layers \( j \) and \( j-1 \) and corresponding bias \( b^{(j)} \) are learned during training [71].

![DNN Architecture Diagram](image)

Figure 4-13 DNN architecture representing \( L \) hidden layers \((h)\), input and output layers are visible and based on the size of the features and the number of classes. The input neurons are weighted, \( H^{(L)} \), by vectors to generate visible, \( v \), outputs. The DNN structure is similar to the shallow NN however contains multiple hidden layers, \( h^{(L+1)} \).

Theoretical benefits of deep learning have been known for a long time. However, training deep networks has been problematic without a learning algorithm capable of finding all layer weights [89]. Poorly random initialized weights in fully supervised back-propagation with gradient descent fail in high generalization deep learning [170][171].
Issues related to deep learning are summarized as follows:

- Requires extensively large amount of labeled training data
- Very slow convergence for multiple hidden layer networks [89]
- Problem solving highly non-convex optimization converging to local optima if randomly initialized weights are too large [112][113][171]
- For small initial weights the gradients in early layers are small so it’s infeasible to train many hidden layers [112].
- Sigmoid activation functions can be problematic in training DNN [89]
- Stochastic gradient descent can be unstable and a vanishing gradient [170]

There has been much research into improving deep networks as summarized in [219] and is facilitated by deep learning methods [220][71]. Hinton et al., proposed a new initialization method using unsupervised pre-training [72][71]. Replacing random initialization improved convergence [220][71][112][227].

Each DBN layer is independently pre-trained with unsupervised learning as a 2 layer restricted Boltzmann machines (RBM) [114]. This is shown by Figure 4-14, where a $N$-layer DBN is built up of $N$ stacks of RBM. The RBM are trained in a greedy layer-wise approach as a generative stochastic network [71].

Figure 4-14 Illustration of DBN layer interconnections of the weight matrix, $W$, with visible layer units, $v_m$, and the hidden units, $h_n$ (left) and representation of the layer-wise training of a 3-layer DBN as an unsupervised RBM (right). Where the input to each RBM layer, $v$, is the output of the previous with hidden layers denoted by $h$. 
The input layer to the first RBM corresponds to the visible units, \( v \), from the training data and the first hidden layer of the DBN. During pre-training the RBM parameters are learned as hidden units and become feature detectors for training the next RBM in the stack. The second RBM corresponds to the first and second hidden layers of the DBN and the stacking continues for the following DBN layers.

The energy of a given joint configuration \((i,v)\) of the RBM is described by the energy function, \( E(v,h) \), in equation (4.39). Where \( W=(w_{i,j}) \) is the weight matrix associated with the connection between hidden unit \( h_j \) and visible unit \( v_i \). The bias values for the visible and hidden units are respectively given by \( a_i \) and \( b_j \) [112].

\[
E(v,h) = - \sum_{i \in \text{pixels}} a_i v_i - \sum_{i \in \text{features}} b_j h_j - \sum_i \sum_j v_i h_j w_{ij}
\]  

(4.39)

The probabilities of the states a RBM can take is represented by the probability distribution over the hidden and visible vectors, given by equation (4.40).

\[
P(v,h) = \frac{\sum_h e^{-E(v,h)}}{\sum_v \sum_h e^{-E(v,h)}}
\]  

(4.40)

The RBM has no intra-layer connections so the hidden unit activations are mutually independent given the visible unit activations and vise versa [162]. Therefore for \( m \) visible units and \( n \) hidden units, the conditional probability of a configuration of visible units \( v \), given a configuration of the hidden units \( h \) is given by equation (4.41). Similar, the conditional probability of \( h \) given \( v \) is defined by equation (4.42), where \( v_i \) is the state of pixel \( i \) and \( h_j \) is the state of the hidden unit \( j \).
\[ P(v|h) = \prod_{i=1}^{m} P(v_i|h) \]  \tag{4.41}

\[ P(h|v) = \prod_{j=1}^{n} P(h_j|v) \]  \tag{4.42}

It is common to restrict the RBM to binary units so \( v_i \) and \( h_i \in \{0,1\} \) and the probabilistic version of the neuron activation function is given by equations (4.43) and (4.44). Where \( \sigma \) is the logistic sigmoid function, equation (4.45), with outputs between 0 and 1. The weight between \( i \) and \( j \) is \( w_{ij} \) and \( b_j \) and \( a_i \) are biases.

\[ P(h_j = 1|v) = \sigma \left( b_j + \sum_{i=1}^{m} w_{ji} v_i \right) \]  \tag{4.43}

\[ P(v_i = 1|h) = \sigma \left( a_i + \sum_{j=1}^{n} w_{ji} h_j \right) \]  \tag{4.44}

\[ \sigma(x) = \frac{1}{1 + e^{-x}} \]  \tag{4.45}

The training algorithm learns RBM parameters (\( w_{ij}, a_i \) and \( b_j \)) that maximize the product of the probabilities assigned to the feature vector \( v \) given by equation (4.46) and simplified to the expected log probability given by equation (4.47).

\[ \text{argmax}_{W} \prod_{v \in V} P(v) \]  \tag{4.46}

\[ \text{argmax}_{W} E \left[ \sum_{v \in V} log P(v) \right] \]  \tag{4.47}
The RBM learns using unsupervised training and propagating stochastically through the RBM. Training to optimize the weights is traditionally computed using contrastive divergence algorithm [114]. Contrastive-Divergence updates the weights using Gibbs sampling of the probabilities in equations (4.43) and (4.44), within a gradient descent algorithm [346][211].

\[ \Delta w_{ij} = \varepsilon \left( \langle v_i h_j \rangle_{\text{data}} - \langle v_i h_j \rangle_{\text{recon}} \right) \]  (4.48)

The weights, \( w_{ij} \), associated with connections between the \( h_j \) and \( v_i \) units are updated by equation (4.48); where \( \varepsilon \) is the learning rate. The positive gradient, \( \langle v_i h_j \rangle_{\text{data}} \), is represented as the outer product of visible and hidden units. Likewise, the negative gradient, \( \langle v_i h_j \rangle_{\text{recon}} \), is the outer product of the reconstructed units (from Gibbs sampling). The bias updates, \( \Delta a_i \) and \( \Delta b_i \), are explained analogously.

\[ \Delta a_i = \varepsilon \left( \langle v_i \rangle_{\text{data}} - \langle v_i \rangle_{\text{recon}} \right) \]  (4.49)

\[ \Delta b_j = \varepsilon \left( \langle h_j \rangle_{\text{data}} - \langle h_j \rangle_{\text{recon}} \right) \]  (4.50)

After unsupervised pre-training, with the RBM algorithm, a good initialization for the hidden weights and biases is obtained. The output layer is randomly initialized and the entire DNN is fine tuned with supervised learning via back-propagation and gradient descent, which works well if the initial weights are close to a good solution.

Hyper parameters including mini-batch size, learning rate and epochs are not found during the learning algorithm yet affect DNN performance. There is a benefit of heuristics to optimize hyper-parameters but is very time consuming. Instead hyper-parameters are determined with trial and error and based on cross-validated accuracy.
4.7 New Optimized Multichannel Weighted Deep Neural Network (OMC-DNN)

Certain features may be weak on their own but combined with more features performance can be improved. This can be achieved using early fusion (feature fusion) or late-integration (decision and score fusion).

Early fusion is simple to implement but has disadvantages with large-dimension fused features causing training computational difficulties due to the “curse of dimensionality”. Also in feature fusion noisy or poor features can compromise reliable features resulting in a fusion that can be worse than the individual features.

Late-integration allows takes into account the reliability of each feature with adaptive weighting to optimize features in combination. Therefore, in this case the features were fused at a decision level using multiple machine learning models.

In contrast to single-channel classification, with a single feature set, multi-channel classification uses parallel single channel classifiers. In this application multiple single channel DNNs (SC-DNN), as described in Section 4.6.4, were trained independently with distinctive feature sets with the results fused a decision level with weights.

The weights were determined in an automatic optimization procedure using a hybrid Simulated Annealing/Genetic Algorithm (SA/GA), described in Section 4.7.3. The following sub-sections explain the optimized multichannel weighted DNN (OMC-DNN) methodology and optimization procedure.
4.7.1 Optimized Multichannel DNN (OMC-DNN) system description

The $K$ channel OMC-DNN given by Figure 4-15 classifies an unknown test sample with the following two stages:

**Stage 1:** Unknown test samples are classified using parallel single-channel DNN classifiers with different feature sets. This provides independent class estimates, $\hat{y}_k(x_i), k = 1, ..., K$, for $K$ feature channels, for a given speech sample $x_i$.

**Stage 2:** The SC-DNN results from stage 1 are combined using a weighted sum at the decision level, given by equation (4.51), for a final decision. The weights are pre-determined during an optimization phase using hybrid SA/GA.

![Diagram of OMC-DNN system](image)

Figure 4-15 Process of testing unknown sample, $x_i$, with the general $K$ channel OMC-DNN such that $K$ independent feature sets are classified by individual SC-DNN with the results, $\hat{y}_k(x_i)$, weighted by optimized weights, $w_k^{\text{opt}}$, and the final class result, $\hat{h}(x_i)$, converted by the decision matrix.
In more detail, each channel trained an independent SC-DNN reducing their MSE objective function resulting in class estimates, \( \hat{y}_k(x_i) \), from each channel \( k = 1, 2, ..., K \). Then using late integration a combined decision, \( s(x_i) \), was defined as the weighted sum of each channel fused at a decision level, given by equation (4.51).

\[
s(x_i) = \sum_{k=1}^{K} w_k^{\text{opt}} \hat{y}_k (x_i)
\]  

(4.51)

Where \( i \) is the sample number, \( \hat{y}_k(x_i) \) is the estimate of channel \( k \), and \( w_k^{\text{opt}} \) denotes the optimized weights. The procedure to find the weights is described later in Section 4.7.2 using the optimization algorithm described in Section 4.7.3.

Table 4-2 shows the decision matrix used the sign of \( s(x_i) \) to map the final decision, \( \hat{h}(x_i) \), corresponding to a predicted binary emotion. Similarly, for the multiclass case Table 4-3 defined the decision matrix where \( \hat{h}(x_i) \) was selected based on a criteria range of \( s(x_i) \). The ranges of \( s(x_i) \) associated with each multiclass emotion were experimentally determined through trial and error established on increasing overall OMC-DNN accuracy.

**Table 4-2 Binary OMC-DNN Decision Matrix to Convert the Weighted Result, \( s(x_i) \), to Determine the Final Result, \( \hat{h}(x_i) \), and the Corresponding Predicted Class (Emotion or Neutral)**

<table>
<thead>
<tr>
<th>Sign of ( s(x_i) )</th>
<th>Value of ( \hat{h}(x_i) )</th>
<th>Predicted Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>+1</td>
<td>Emotion</td>
</tr>
<tr>
<td>-</td>
<td>-1</td>
<td>Neutral</td>
</tr>
</tbody>
</table>

**Table 4-3 Multiclass OMC-DNN Decision Matrix to Convert the Weighted Result, \( s(x_i) \), to Determine the Final Result, \( \hat{h}(x_i) \), and the Corresponding Predicted Class**

<table>
<thead>
<tr>
<th>Range of ( s(x_i) )</th>
<th>Value of ( \hat{h}(x_i) )</th>
<th>Predicted Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>( s&lt;1.4 )</td>
<td>1</td>
<td>Angry (A)</td>
</tr>
<tr>
<td>1.5&lt;( s \leq 2.4 )</td>
<td>2</td>
<td>Bored (B)</td>
</tr>
<tr>
<td>2.5&lt;( s \leq 3.4 )</td>
<td>3</td>
<td>Neutral (N)</td>
</tr>
<tr>
<td>3.5&lt;( s \leq 4.4 )</td>
<td>4</td>
<td>Disgust (D)</td>
</tr>
<tr>
<td>4.5&lt;( s \leq 5.4 )</td>
<td>5</td>
<td>Sad (S)</td>
</tr>
<tr>
<td>5.5&lt;( s \leq 6.4 )</td>
<td>6</td>
<td>Happy (H)</td>
</tr>
<tr>
<td>6.5&lt;( s )</td>
<td>7</td>
<td>Fear (F)</td>
</tr>
</tbody>
</table>
4.7.2 Determining the Optimal weight

Late-integration takes into account the reliability of each feature with the results fused at a weighted decision-level. Under certain conditions, the optimal weights are inversely proportional to the single channel’s normalized objective function [92][93][91][90]. In these cases the importance of each feature is ranked on individual performance, not how they would perform combined. Instead it has been shown globally optimized weights improve performance of multichannel systems [163][92].

Figure 4-16 Procedure for the general K channel OMC-DNN system to determine the optimized set of channel weights \( w_k^{\text{opt}} \) using the Simulated Annealing Genetic Algorithm Hybrid (SA/GA) optimization algorithm. The weights are iteratively updated, \( w_k^{\text{update}} \), until convergence based on the objective function, \( f_{\text{obj}}(w) \) that increases OMC-DNN accuracy.
The optimized weights were determined automatically during a training phase, using the classification system and optimization algorithm as shown by Figure 4-16. Independent SC-DNN were trained in supervised classification so each sample \( x_i \) \((i=1, \ldots, N)\) has a known class \( y(x_i) \), required to compare performance of weight sets.

The initial weights are equal for each feature channel such that, \( W_{k}^{\text{init}} = \frac{1}{K} \). For each sample, \( x_i \), the current weight set \( W^\text{update} = \{ w_{1}^\text{update}, \ldots, w_{K}^\text{update} \} \) and channel outputs, \( y_{k}(x_i), k=1, \ldots, K \), determine the output, \( s(x_i) \), of the OMC-DNN is defined by (4.52), and converted to the class, \( h(x_i) \), with the decision matrix.

\[
 s(x_i) = \sum_{k=1}^{K} w_{k} y_{k}(x_i) \tag{4.52}
\]

The weights were iteratively updated until the convergence of the objective function \( f_{\text{obj}}(w) \) and an optimal set of weights. The objective function minimizes the MSE between the actual class, \( y(x_i) \), and the class estimate, \( \hat{h}(x_i) \), defined by equation (4.53). So the optimal weights, \( W^{\text{opt}} = \{ W_{1}^{\text{opt}}, \ldots, W_{M}^{\text{opt}} \} \), lead to the best accuracy in the overall OMC-DNN system impartial to individual performance.

\[
 f_{\text{obj}}(w) = \frac{1}{N} \sum_{i=1}^{N} \left( y(x_i) - \hat{h}(x_i) \right)^2 \tag{4.53}
\]

4.7.3 Optimization algorithm

The optimal weights can be found with an exhaustive or grid search but are both inefficient. It is ideal to use an efficient algorithm, such as iterative, heuristic or optimization, to find weights. Many algorithms often get stuck in local minima and struggle to reach a global solution. In application the optimization algorithm was a hybrid of two meta-heuristic search strategies: Simulated annealing (SA) [119] and a genetic algorithm (GA) [308] as described in the following sub-sections.
4.7.3.1 Simulated Annealing

The Metropolis algorithm (MA) is based on the law of thermodynamics that states at temperature $T$, the probability of a change in energy, $\Delta E$, is given by Boltzmann’s probability function in equation (4.54) where $k$ is the Boltzmann constant.

MA calculates the cost of a solution by the energy change, $\Delta E = E(x') - E(x_0)$, of the current solution, $x_0$, and a new randomly generated solution $x'$. A decrease in energy is a good trade and the solution is accepted.

$$p = e^{-\frac{\Delta E}{kT}}$$

(4.54)

$$\zeta < p$$

(4.55)

Some solutions that increase the cost function are accepted to explore larger solution space to not get stuck in local minima [120]. The Boltzmann acceptance criteria, equation (4.55) where $0<\zeta<1$, decides if a worse solution is accepted. Iterations are continued until the optimal solution is found or criteria are met.

MA has met issues in finding optimal solution at the bottom of a “long narrow valley” with an inability to explore local downhill movements and has low search efficiency and a large convergence time. Kirkpatrick et al. overcame these issues by applying the MA with SA to search feasible solutions [119].

SA mimics the evolution of a solid heated, during temperature reduction to thermal equilibrium. MA is repeated as the temperature, $T_i$, reduces until the system freezes given by Figure 4-17.
Figure 4-17 Flow diagram outlining the process of Simulated Annealing (SA) including the criteria for acceptance of worse solutions and stopping criteria

The probability of accepting a worse solution is now a function of both the change in the energy (cost function) and temperature, shown by Figure 4-18. Initially, the algorithm searches a wide range and accepts solutions with increased energy and as the temperature decreases the probability of accepting a worse move is reduced.

Figure 4-18 The Boltzmann’s acceptance probability function used in the classical Metropolis algorithm and simulated annealing. The temperature reduces with iteration number where $T_1 > T_2$, the threshold, $\zeta$, is a random number between 0 and 1.
The cooling schedule of SA has four components: initial temperature, $T_0$, final temperature, $T_f$, temperature decrement and iterations at each temperature. Many methods can control the cooling schedule and in this case temperature is updated using equations (4.56) and (4.57), where $I_{T_{\text{max}}}$ is the maximum temperature change

$$T = T_0 e^{-T_{\text{chain}} \alpha}$$

(4.56)

$$\alpha = -\left(\log \left(\frac{T_f}{T_0}\right)\right) \frac{I_{T_{\text{max}}}}{I_{T_{\text{max}}}}$$

(4.57)

4.7.3.2 Hybrid: Simulated Annealing with Genetic Algorithm (SA/GA)

Hybridization combines search algorithms to optimize solutions and improve the convergence [166] and improves optimization in many applications than a single optimization algorithm [165][164].

Processes that simulate natural selection and evolution are generally used for hybridization. Evolutionary search methods, such as genetic algorithms, evolution strategies, and genetic programming, are advantageous due to simplicity, robustness and flexibility [155].

The genetic aspect of the algorithm generates a population of possible solution. For each population member the objective function of the optimization algorithm is evaluated and the best solutions preserved as offspring populations for the succeeding iterations. Local minima are avoided by occasionally allowing a mutation process; this creates a random perturbation to all of the population members.

The hybrid SA/GA algorithm has two stages; first the GA generates a population of initial solutions and looks at these in a global sense. Second, SA optimizes each individual solution locally and the best solutions are used in the next generation of the GA and repeats until the specified criteria are met.
4.8 Acoustic Emotion Recognition Modeling and Classification Setup

In order to determine the dependability and consistency of using the acoustic features, as images and vectors, towards AER multiple classification setups were explored. The following experiments were based on the database, features and classification system methods in the previous sections. The three major classification categories are investigated with increasing complexity as follows:

1) Single-channel emotion recognition using 1D features with SVM, NN, GMM;
2) Single-channel emotion recognition using 2D features with NN, DNN, DNN-NP
3) New optimized multi-channel DNN emotion recognition using 2D features

Specific details of each of these major setups are described in Sections 4.8.1 and 4.8.2 for single and multichannel and experiments then outlined in Section 4.9. At this stage it should be noted that the experimental configurations are not completely comparable, due to different training/validation/test segmentations and CV runs.

4.8.1 Single channel classifier setups

- **Gaussian Mixture Model (GMM-1D)**

Binary and multiclass GMM was implemented with 1D feature vectors and 5 mixtures, described in Section 4.6.1, using the HTK toolbox [433]. The experimental procedure segments data into 80% train and 20% test sets and 5 fold CV.

- **Support Vector Machine (SVM-1D)**

The SVM was implemented from the MATLAB Optimization Toolbox, outlined in Section 4.6.2, using 1D acoustic features limited to binary AER. The SVM was trained with linear kernel function and the parameters and learned through sequential minimization optimization with least squares as the objective function. The setups involves k-fold CV with 5 folds using segmented training and testing data (80:20).
• **Optimized SVM (LIBSVM-1D)**

LIBSVM C++ implementation was used as a faster alternative with more memory, than the MATLAB toolbox [291], trained using sequential minimization optimization (SMO) [292]. A radial basis function (RBF) was used as the kernel popular in SVM, because of their localized and finite responses across the entire range [279].

The aim was to increase classification accuracy by optimizing hyper-parameters for the cost/penalty, $C$, and the kernel parameter, $\gamma$, using a grid search on the training dataset as suggested by [291]. This process used 3-stages with big, medium and small scales so the search space is reduced with fine-tuning. For each combination of parameters, $C$ and $\gamma$, the SVM was trained with 5-fold CV with 10% of the database.

The remaining 90% of the database was used to tune the $w$ and $b$ parameters of the SVM with the optimal cross-validated $C$ and $\gamma$. The SVM was trained using 5-fold CV with the dataset separated into 80% for training and 20% for the out of sample test set.

• **Neural Network (NN-1D) and (NN-2D)**

The Neural Network (NN), outlined in Section 4.6.3, implemented for 2D images has 1156 input neuron vector. The NN with the standard feature vector the input size is 12 (MFCC) or 17 (TEO). In both options, there was one hidden layer with 50 nodes with a tan-sigmoid activation function with either two (binary) or seven (multiclass) outputs neurons and a linear output function. Training was implemented using Levenberg–Marquardt algorithm with a maximum epoch of 1000. The data was segmented into training, validation and test sets of 70%, 15% and 15% for early stop cross validation.
• **Deep Neural Network (DNN-2D) and (DNN-NP-2D)**

The DNN follows the methodology outlined in Section 4.6.4 with supervised and unsupervised training and DNN-NP without unsupervised pre-training with 2D image features. The input layer to the DNN is 1156 neurons (number of pixels in the images), and has two hidden layers with 100 nodes each. The relatively small size (two 100 node layers) of the DNN was chosen in consideration, especially due to the small dataset, to lower the risk of over-fitting.

The activation functions in the hidden and output layers are both sigmoid. The binary case has two output neurons and seven in the multiclass (representing each emotion). In this setup the DNN has 5 epochs, 30 epochs in the unfolded NN, a 0.35 learning rate and a batch size of 100. The data was split into 80%/20% for training and testing with 3-fold CV.

**4.8.2 Multichannel channel classifiers setups**

The OMC-DNN, outlined in Section 4.7, pre-learned the optimized weights for the decision-fusion of each individual feature channels using 80% of the entire database. This phase segmented the data into subsets of 80% training and 20% testing in 3-fold CV such that the optimized weights increased the overall OMC-DNN accuracy.

The remaining 20% of the database was used to train and test the entire OMC-DNN, using the learned optimized channel weights, with 3-fold CV using subsets of 80% and 20% for training and testing to learn the model parameters.
4.9 Experimental setups

A series of experiments were conducted for single channel AER using the database described in Section 3.4, features from Section 4.4, classifiers from Section 4.6 and setups in Section 4.8.1 outlined as follows:

- **EXP1**: Examined Deep Neural Networks for binary emotion recognition using 2D images of MFCC and TEO feature contours.

- **EXP2**: Investigated the usefulness of MFCC and TEO features extracted from the glottal waveform for AER in comparison to the speech waveform (EXP1).

- **EXP3**: Using features from EXP1-EXP2, the effect of gender was examined comparing gender independent (GIM) and dependent (GDM) models.

- **EXP4**: Repeated previous experiments (EXP1-EXP3) for multiclass AER using 7 simultaneous emotions.

- **EXP5**: The previous experiments (EXP1-EXP4) were repeated with benchmark classifiers including SVM, NN, GMM with 1D or 2D features.

Further investigation into multi-channel AER was examined using an OMC-DNN, as described previously in Section 4.7, based on the setup described in Section 4.8.2 with the following experiments:

- **EXP6**: The DNN performed the best compared to all other classifiers and was next used for a new optimized multichannel DNN. 3-channel and 4-channel OMC-DNN were tested using a combinations of glottal (G-TEO and G-MFCC) and speech (S-MFCC and S-TEO) waveform features and derivatives.
4.10 Evaluation methods

As often used in binary classification assessment [167], binary AER was evaluated by sensitivity, specificity and accuracy given by equations (4.58), (4.59) and (4.60).

\[
\text{Sensitivity} = \frac{TP}{TP + FN} \times 100\% \quad (4.58)
\]

\[
\text{Specificity} = \frac{TN}{TN + FP} \times 100\% \quad (4.59)
\]

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (4.60)
\]

The true positive (TP), false positive (FP), false negative (FN) and true negative (TN) parameters were calculated as defined by the following:

- TP: total number of samples correctly classified as the emotion state
- FP: total number of neutral state samples misclassified as the emotion state
- TN: total number of samples correctly classified as neutral
- FN: total number of emotion state samples misclassified as neutral

For the multiclass case (7 emotions simultaneously) the performance was measured by accuracy, defined as percentage of the correct number of classified emotions of the entire test set, given by equation (4.61). Additionally, the confusion matrix was used to observe how the emotions are misinterpreted.

\[
\text{Accuracy (\%)} = \frac{\text{Number of correctly classified samples}}{\text{Total number of samples}} \times 100\% \quad (4.61)
\]
4.11 Single Channel Emotion Recognition Results

4.11.1 Binary AER using SC-DNN with MFCC and TEO features (EXP1)

The performance of six binary, emotions versus neutral, SC-DNN is given in Table 4-4 using MFCC and TEO. MFCC are stronger for all binary pairs with 85% average accuracy, 14% higher than TEO. For MFCC it was found that anger, joy and disgust perform the best and boredom, fear and sadness worse. This suggests MFCC are good at classifying high arousal emotions against neutral, but struggle discerning between passive emotions. On the other hand TEO performed similar for all binary tests.

<table>
<thead>
<tr>
<th>Emotion</th>
<th>MFCC</th>
<th>TEO</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>94.7</td>
<td>93.6</td>
</tr>
<tr>
<td>B</td>
<td>81.4</td>
<td>73.9</td>
</tr>
<tr>
<td>D</td>
<td>86.1</td>
<td>87.7</td>
</tr>
<tr>
<td>F</td>
<td>72.4</td>
<td>83.1</td>
</tr>
<tr>
<td>J</td>
<td>83.3</td>
<td>73.7</td>
</tr>
<tr>
<td>S</td>
<td>78.4</td>
<td>82.2</td>
</tr>
<tr>
<td>Avg</td>
<td>82.7</td>
<td>86.3</td>
</tr>
</tbody>
</table>

4.11.2 Binary AER using SC-DNN with G-MFCC and G-TEO features (EXP2)

Previous studies found acoustic features derived from the glottal waveform improve classification tasks [65]. Therefore, EXP1 was repeated by extracting MFCC and TEO from the glottal waveform (G-MFCC and G-TEO). The results, in Table 4-5 results, indicate for all binary setups G-MFCC is better than G-TEO on average by 6%. For both features anger and joy are the easiest and boredom the most difficult to detect.

<table>
<thead>
<tr>
<th>Emotion</th>
<th>G-MFCC</th>
<th>G-TEO</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>91.2</td>
<td>86.1</td>
</tr>
<tr>
<td>B</td>
<td>67.4</td>
<td>55.8</td>
</tr>
<tr>
<td>D</td>
<td>76.2</td>
<td>68.7</td>
</tr>
<tr>
<td>F</td>
<td>76.3</td>
<td>68.1</td>
</tr>
<tr>
<td>J</td>
<td>85.8</td>
<td>69.5</td>
</tr>
<tr>
<td>S</td>
<td>77.0</td>
<td>61.1</td>
</tr>
<tr>
<td>Avg</td>
<td>79.0</td>
<td>68.2</td>
</tr>
</tbody>
</table>

109
4.11.3 Comparison of Glottal and Speech Features using DNN for Binary AER

The accuracy of all binary AER setups are given in Figure 4-19, comparing the four features (S-MFCC, G-MFCC, S-TEO, G-TEO). In terms of glottal waveform features there was a 3.8% average accuracy reduction using G-MFCC compared to S-MFCC. A paired t-test, on binary accuracies, showed S-MFCC was significantly better ($p=0.018$) than G-MFCC. In contrast, there was a boost in accuracy for G-TEO by an average of 4.8% compared to S-TEO and is statistically significant $p=0.048$.

This improvement provides evidence supporting the fact glottal waveform is effective in improving speech recognition applications [65]. This is true in the case of G-TEO, but not with respect to G-MFCC compared to S-MFCC. This could be due to G-TEO capacity to represent the non-linearity of airflow of the glottal cycle [65][68].
4.11.4 Effectiveness of GDMs and GIM for AER using SC-DNN (EXP3)

It is known acoustic features differ between males and females, especially during emotional speech [235][234][172]. An experiment into the effectiveness of Gender Dependent Models (GDMs) was assessed using S-MFCC, S-TEO, G-MFCC and G-TEO. The GDM DNN were trained and tested with male (GDM-M) or female (GDM-F) subjects and the Gender Independent Model (GIM) with both genders.

Table 4-6 Average Accuracy, Sensitivity and Specificity of Binary AER using a SC-DNN comparing MFCC and TEO features with both Gender Dependent Models (GDM-M and GDM-F)

<table>
<thead>
<tr>
<th></th>
<th>GDM-M</th>
<th>GDM-F</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MFCC</td>
<td>TEO</td>
</tr>
<tr>
<td></td>
<td>Sens</td>
<td>Spec</td>
</tr>
<tr>
<td>A</td>
<td>97.3</td>
<td>95.0</td>
</tr>
<tr>
<td>B</td>
<td>82.3</td>
<td>75.7</td>
</tr>
<tr>
<td>D</td>
<td>91.2</td>
<td>94.1</td>
</tr>
<tr>
<td>F</td>
<td>83.7</td>
<td>93.2</td>
</tr>
<tr>
<td>J</td>
<td>83.7</td>
<td>96.7</td>
</tr>
<tr>
<td>S</td>
<td>83.8</td>
<td>89.1</td>
</tr>
<tr>
<td>Avg</td>
<td>87.0</td>
<td>90.6</td>
</tr>
</tbody>
</table>

Table 4-7 Average Accuracy, Sensitivity and Specificity of Binary AER using a SC-DNN comparing G-MFCC and G-TEO features with both Gender Dependent Models (GDM-M and GDM-F)

<table>
<thead>
<tr>
<th></th>
<th>GDM-M</th>
<th>GDM-F</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>G-MFCC</td>
<td>G-TEO</td>
</tr>
<tr>
<td></td>
<td>Sens</td>
<td>Spec</td>
</tr>
<tr>
<td>A</td>
<td>90.8</td>
<td>96.3</td>
</tr>
<tr>
<td>B</td>
<td>73.8</td>
<td>60.4</td>
</tr>
<tr>
<td>D</td>
<td>88.1</td>
<td>87.8</td>
</tr>
<tr>
<td>F</td>
<td>76.1</td>
<td>76.7</td>
</tr>
<tr>
<td>J</td>
<td>85.1</td>
<td>95.4</td>
</tr>
<tr>
<td>S</td>
<td>75.2</td>
<td>78.6</td>
</tr>
<tr>
<td>Avg</td>
<td>81.5</td>
<td>82.5</td>
</tr>
</tbody>
</table>
Table 4-6 and Table 4-7 give the performance of binary AER of GDMs using speech and glottal waveform features respectively. Overall TEO are the worst and MFCC the best feature by an average of 11% and 13% for GDM-M and GDM-F.

Based on a paired t-test it was found that the accuracy of MFCC was significantly better compared to G-MFCC for GDM-M ($p=0.008$) and GDM-F ($p=0.035$) by an average of 9% and 5%. There was improvement using G-TEO compared to TEO by an average of 6% and 8% for GDM-F and GDM-M; but was not significant with $p=0.06$ (GDM-F) and $p=0.0157$ (GDM-M).

Generally GDM-M performed best with anger and happiness, whilst emotions such as boredom, sadness or fear performed worse. The GDM-F had anger, happiness and fear as the top emotions and boredom the least detectable emotion against neutral.

![Figure 4-20 Comparison of accuracy between gender independent (GIM) and gender dependent models (GDM) using DNN for the 6 different binary emotion recognition tasks. Results are supplied separately for the MFCC, TEO, G-MFCC and G-TEO.](image-url)
Figure 4-20 directly compares binary classification accuracy for GIM and GDMs using S-MFCC, G-MFCC, S-TEO and G-TEO. It was observed the GIM accuracy was lower than GDM for all features and emotions. GDM-F displayed a clear advantage in fear detection compared to the GDM-M for all features.

To a lesser extent, boredom had higher accuracy for GDM-M compared to GDM-F. This implies gender plays a role in emotion recognition depending on the emotion and feature. MFCC could be considered more robust than TEO to gender differences as the range of results of GIM and GDMs are closer with MFCC compared to the range present with TEO.

Table 4-8 compares the results, comparing gender models, with the average accuracy of binary AER. Compared to GIM on average GDM are more accurate by approximately 4% 0% 8% 10% for GDM-M and 4% 3% 5% 9% for the GDM-F for S-MFCC, G-MFCC, S-TEO and G-TEO features.
4.11.5 GDMs and GIM for Multiclass AER using SC-DNN (EXP4)

EXP1-EXP3 were repeated for multiclass AER using the same acoustic features and DNN with seven outputs. The GIM confusion matrices are supplied to provide strengths, weaknesses and overall multiclass accuracy given in Table 4-9 and Table 4-10 for speech (S-MFCC, S-TEO) and glottal (G-MFCC, G-TEO) features.

Table 4-9 CONFUSION MATRIX FOR GIM MULTICLASS AER WITH SC-DNN USING S-MFCC AND S-TEO FEATURES (A: ANGER, B: BORED, D: DISGUST, F: FEAR, J: JOY AND S: SAD)

<table>
<thead>
<tr>
<th></th>
<th>S-MFCC</th>
<th></th>
<th>S-TEO</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
<td>D</td>
</tr>
<tr>
<td>A</td>
<td>80</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>B</td>
<td>3</td>
<td>64</td>
<td>6</td>
</tr>
<tr>
<td>D</td>
<td>10</td>
<td>9</td>
<td>55</td>
</tr>
<tr>
<td>F</td>
<td>14</td>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td>J</td>
<td>24</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>N</td>
<td>3</td>
<td>24</td>
<td>7</td>
</tr>
<tr>
<td>S</td>
<td>2</td>
<td>16</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 4-10 CONFUSION MATRIX FOR GIM MULTICLASS AER WITH SC-DNN USING G-MFCC AND G-TEO FEATURES (A: ANGER, B: BORED, D: DISGUST, F: FEAR, J: JOY AND S: SAD)

<table>
<thead>
<tr>
<th></th>
<th>G-MFCC</th>
<th></th>
<th>G-TEO</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
<td>D</td>
</tr>
<tr>
<td>A</td>
<td>87</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>9</td>
<td>50</td>
<td>4</td>
</tr>
<tr>
<td>D</td>
<td>23</td>
<td>16</td>
<td>36</td>
</tr>
<tr>
<td>F</td>
<td>27</td>
<td>20</td>
<td>8</td>
</tr>
<tr>
<td>J</td>
<td>45</td>
<td>11</td>
<td>8</td>
</tr>
<tr>
<td>N</td>
<td>9</td>
<td>32</td>
<td>5</td>
</tr>
<tr>
<td>S</td>
<td>3</td>
<td>20</td>
<td>2</td>
</tr>
</tbody>
</table>

The S-MFCC, G-MFCC and G-TEO confusion matrices show anger was the easiest emotion to detect and fear, joy, disgust and neutral the hardest. Joy was the easiest and sadness the most difficult with S-TEO. The results reveal anger was the most detectable emotion, which is high arousal emotion compared to passive (neutral). Joy is often misclassified as anger and similar issues are evident between passive emotions of sadness, boredom and neutral. This was expected given the spectral similarities of these emotions and observations from past studies [178][356].
Table 4-11 **Average Accuracy (%) of Multiclass (7-Classes) SC-DNN Using the S-MFCC, S-TEO, G-MFCC and G-TEO Features with Gender Dependence (GDM-M and GDM-F) and Gender Independence (GIM)**

<table>
<thead>
<tr>
<th></th>
<th>S-MFCC</th>
<th>G-MFCC</th>
<th>S-TEO</th>
<th>G-TEO</th>
</tr>
</thead>
<tbody>
<tr>
<td>GIM</td>
<td>58</td>
<td>46</td>
<td>27</td>
<td>35</td>
</tr>
<tr>
<td>GDM-M</td>
<td>61</td>
<td>53</td>
<td>33</td>
<td>35</td>
</tr>
<tr>
<td>GDM-F</td>
<td>60</td>
<td>52</td>
<td>38</td>
<td>40</td>
</tr>
</tbody>
</table>

The multiclass AER accuracies for S-MFCC, G-MFCC, S-TEO and G-TEO with GIM and GDMs are shown in Table 4-11. As with binary AER, S-MFCC is the best and S-TEO was the worst for all multiclass models (GDMs and GIM).

Additionally, extracting TEO from the glottal waveform (G-TEO) improved compared to speech waveform (S-TEO) by 9%, 2% and 2% for GIM, GDM-M and GDM-F respectively. Although, MFCC derived from the glottal waveform (G-MFCC) reduced the accuracy of S-MFCC by 12%, 7% and 8%.

The GDM-M improved the GIM by 3%, 7%, 6% and 0% for S-MFCC, G-MFCC, S-TEO and G-TEO. Similar for the GDM-F the multiclass accuracy improved by 2%, 6%, 11% and 5% compared to the GIM.

Considering seven classes are simultaneously classified the possibility of randomly selecting the correct class is 14%. All of the multiclass models are above this chance level in overall accuracy and individual emotions. The highest accuracy achieved in 7-class AER was 61% using the S-MFCC with a GDM-F.
4.11.6 Comparison of emotion recognition with benchmark classifiers (EXP5)

The DNN was compared with benchmark classifiers, outlined in Sections 4.6 and 4.8, by the average binary accuracy in Table 4-12 for six classifiers, four features and three gender models. The performance was consistent with all classifiers: MFCC the best, TEO the worst, G-TEO better compared to S-TEO and GDM improves on GIM.

DNN with unsupervised pre-training (DNN-2D) improved compared to no pre-training (DNN-2D-NP) by 5.25% (GIM), 10% (GDM-F) and 10.25% (GDM-M). NN using proposed 2D features (NN-2D) performed worse than feature vectors (NN-1D).

| Table 4-12 AVERAGE ACCURACY (%) OF 6 BINARY AER SYSTEMS WITH A COMPARISON OF DNN AND BENCHMARK CLASSIFIERS (NN, GMM AND SVM) USING S-MFCC, G-MFCC, S-TEO, G-TEO WITH 2D OR 1D FEATURES FOR GDM AND GIM. |
|------------------|------------------|------------------|------------------|------------------|------------------|------------------|
| Classifier/ Feature | DNN-2D | DNN-NP-2D | NN-1D | NN-2D | GMM-1D | SVM | LIBSVM |
|------------------|------------------|------------------|------------------|------------------|------------------|------------------|
| GIM | | | | | | | |
| S-MFCC | 85 | 76 | 82 | 67 | 78 | 75 | 76 |
| G-MFCC | 81 | 78 | 78 | 71 | 71 | 73 | 80 |
| S-TEO | 71 | 64 | 62 | 56 | 61 | 58 | 66 |
| G-TEO | 75 | 73 | 72 | 66 | 69 | 68 | 68 |
| GDM-F | | | | | | | |
| S-MFCC | 89 | 79 | 85 | 75 | 81 | 78 | 84 |
| G-MFCC | 84 | 79 | 84 | 74 | 74 | 76 | 78 |
| S-TEO | 76 | 67 | 63 | 55 | 60 | 61 | 65 |
| G-TEO | 84 | 68 | 72 | 66 | 67 | 69 | 71 |
| GDM-M | | | | | | | |
| S-MFCC | 89 | 78 | 85 | 69 | 81 | 75 | 81 |
| G-MFCC | 81 | 75 | 81 | 74 | 76 | 73 | 81 |
| S-TEO | 79 | 71 | 64 | 58 | 61 | 60 | 71 |
| G-TEO | 85 | 69 | 74 | 68 | 69 | 70 | 72 |

Overall the DNN was best classifier, for all features and gender models, performing on average 15%, 12%, 11%, 8.5%, 7.2% and 6.4% better than the NN-2D, SVM, GMM, DNN-NP, LIBSVM and NN-1D classifiers. The DNN accuracy has statistically significant improvement \((p<0.05)\) compared to benchmark classifiers for most setups; shown by the paired t-test \(p\)-values in Table 4-13. The LIBSVM-1D, with optimized \(C\) and \(\gamma\) parameters, was the only classifier mostly non-significant.
A comparison of multiclass classifiers was restricted to NN, GMM and DNN because SVM is mostly for binary classification and DNN without pre-training showed degradation. The DNN outperformed all classifiers in every gender and feature case. The GIM DNN on average for all features was 7.6% and 5.9% more accurate than the NN and GMM. Furthermore, the NN and GMM are improved by 10.2% and 5.0% for GDM-M and 9.4% and 11.2% for GDM-F using the DNN.

Table 4-14 AVERAGE ACCURACY FOR MULTICLASS AER USING 2D FEATURES IN THE DNN AND FEATURE VECTORS FOR THE NN AND GMM BASELINE CLASSIFIERS. RESULTS ARE SHOWN FOR THE GIM AND GDMs AND FOUR FEATURE SETS

<table>
<thead>
<tr>
<th>Feature/Classifier</th>
<th>DNN-NP-2D</th>
<th>NN-1D</th>
<th>NN-2D</th>
<th>GMM-1D</th>
<th>SVM-1D</th>
<th>LIBSVM-1D</th>
</tr>
</thead>
<tbody>
<tr>
<td>GIM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S-MFCC</td>
<td>0.001</td>
<td>0.008</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>0.001</td>
<td>0.021</td>
</tr>
<tr>
<td>G-MFCC</td>
<td>0.024</td>
<td>0.016</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>0.001</td>
<td>0.699</td>
</tr>
<tr>
<td>S-TEO</td>
<td>0.004</td>
<td>0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>0.002</td>
<td>0.048</td>
</tr>
<tr>
<td>G-TEO</td>
<td>0.039</td>
<td>0.048</td>
<td>0.001</td>
<td>0.003</td>
<td>0.001</td>
<td>0.014</td>
</tr>
<tr>
<td>GDM-M</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S-MFCC</td>
<td>0.008</td>
<td>0.009</td>
<td>0.001</td>
<td>&lt;0.001</td>
<td>0.006</td>
<td>0.197</td>
</tr>
<tr>
<td>G-MFCC</td>
<td>0.002</td>
<td>0.929</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>0.376</td>
</tr>
<tr>
<td>S-TEO</td>
<td>0.007</td>
<td>0.002</td>
<td>&lt;0.001</td>
<td>0.002</td>
<td>0.001</td>
<td>0.007</td>
</tr>
<tr>
<td>G-TEO</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>0.092</td>
</tr>
<tr>
<td>GDM-F</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S-MFCC</td>
<td>&lt;0.001</td>
<td>0.004</td>
<td>&lt;0.001</td>
<td>0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>G-MFCC</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>0.359</td>
</tr>
<tr>
<td>S-TEO</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>0.086</td>
</tr>
<tr>
<td>G-TEO</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>0.142</td>
</tr>
</tbody>
</table>

The results suggest the NN-2D can’t fully utilize the 2D features and the DNN stronger modeling capabilities are required to capture 2D higher-level abstract cues [173]. Improved performance could be that the more complex, spatial structure provides higher compatibility with the multi-layered structure of the DNN.
4.12 Optimized Multichannel DNN (OMC-DNN) emotion recognition results

The results in previous sections showed the proficiency of DNN for both binary and multiclass AER using individual feature sets. Features that performed poorly or well individually could be improved by feature combination by providing valuable information. This was undertaken using late-integration of multiple independent weighted feature channels using an OMC-DNN with the following feature sets:

1) 3 channels:
   a) **S-MFCC set**: Spectral MFCCs, ΔMFCC and ΔΔMFCC
   b) **G-MFCC set**: Glottal MFCCs, ΔMFCC and ΔΔMFCC
   c) **S-TEO set**: Spectral TEOs, ΔTEO and ΔΔTEO
   d) **G-TEO set**: Glottal TEOs, ΔTEO and ΔΔTEO

2) 4 channels:
   a) **Static set**: S-TEO, S-MFCC, G-TEO, G-MFCC
   b) **Derivative set**: ΔS-TEO, ΔS-MFCC, ΔG-TEO, ΔG-MFCC
   c) **Double Derivative set**: ΔΔS-TEO, ΔΔS-MFCC, ΔΔG-TEO, ΔΔG-MFCC

The OMC-DNN follows the methodology described in Section 4.7, with a new two-stage multichannel AER classification. At the first stage, the samples were classified in an independent parallel configuration of different feature channels. At the second stage, the final decision was the weighted sum of the individual channel decisions.
4.12.1 Effectiveness of emotion recognition with OMC-DNN (EXP6)

4.12.1.1 Sets A-D as 3-channel OMC-DNN

OMC-DNN was implemented with four different three-channel feature sets (A, B, C, D) described in Table 4-15. The first channel of each set used a feature form the previous SC-DNN experiments (MFCC, G-MFCC, TEO or G-TEO). The first and second derivatives of the respective feature are used in the second and third channels.

Table 4-15 OUTLINE OF THE FOUR DIFFERENT 3-CHANNEL FEATURE SETS (A-D) FOR THE OMC-DNN EMOTION RECOGNITION EXPERIMENTS

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Channel 1</th>
<th>Channel 2</th>
<th>Channel 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>S-MFCC</td>
<td>ΔS-MFCC</td>
<td>ΔΔS-MFCC</td>
</tr>
<tr>
<td>B</td>
<td>G-MFCC</td>
<td>ΔG-MFCC</td>
<td>ΔΔG-MFCC</td>
</tr>
<tr>
<td>C</td>
<td>S-TEO</td>
<td>ΔS-TEO</td>
<td>ΔΔS-TEO</td>
</tr>
<tr>
<td>D</td>
<td>G-TEO</td>
<td>ΔG-TEO</td>
<td>ΔΔG-TEO</td>
</tr>
</tbody>
</table>

The optimized channel weights given in Table 4-16, averaged for six binary models, represent the importance of each feature channel in the OMC-DNN. Channel weight variation was relatively large in set A, with MFCC more correlated in AER than derivatives. This is true to a lesser extent in set B with G-MFCC. TEO in set C is not always the highest ranked showing importance of the derivatives. Each set D channel weight is similar which indicates G-TEO and derivatives are just as critical.

Table 4-16 AVERAGE WEIGHT ALLOCATION, OF SIX BINARY AER PAIRS, FOR FEATURE SETS A, B, C AND D (STATIC, DERIVATIVE, DOUBLE DERIVATIVE FEATURE CHANNELS) OF THE 3-CHANNEL OMC-DNN IN GIM AND GDM

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Average Weight Allocation of independent feature channels</th>
<th>GIM</th>
<th>GDM-M</th>
<th>GDM-F</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Ch. 1</td>
<td>Ch. 2</td>
<td>Ch. 3</td>
</tr>
<tr>
<td>A</td>
<td></td>
<td>Static</td>
<td>Δ</td>
<td>ΔΔ</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.621</td>
<td>0.252</td>
<td>0.127</td>
</tr>
<tr>
<td>B</td>
<td></td>
<td>0.474</td>
<td>0.268</td>
<td>0.258</td>
</tr>
<tr>
<td>C</td>
<td></td>
<td>0.406</td>
<td>0.326</td>
<td>0.267</td>
</tr>
<tr>
<td>D</td>
<td></td>
<td>0.371</td>
<td>0.327</td>
<td>0.302</td>
</tr>
</tbody>
</table>

Binary OMC-DNN results, given in Table 4-17, for GIM and GDM can be directly compared with the corresponding SC-DNN. In the GIM case sets A and C are comparable to MFCC and TEO in Table 4-4 and sets B and D, for glottal waveform features compared to Table 4-5.
In GDM cases OMC-DNN was compared with SC-DNN in Table 4-6 and Table 4-7 for speech and glottal features. All OMC-DNN setups, with derivatives as additional channels, improved SC-DNN performance.

Table 4-17 AVERAGE ACCURACY (%) FOR THE SIX BINARY AER TESTS USING THE 3-CHANNEL OMC-DNN WITH FOUR DIFFERENT FEATURE SETS (A, B, C AND D) WITH GIM AND GDM (MALE AND FEMALE)

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Multi-Channel Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Anger</td>
</tr>
<tr>
<td>GIM</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>95.8</td>
</tr>
<tr>
<td>B</td>
<td>94.9</td>
</tr>
<tr>
<td>C</td>
<td>82.0</td>
</tr>
<tr>
<td>D</td>
<td>90.3</td>
</tr>
<tr>
<td>GDM-F</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>96.6</td>
</tr>
<tr>
<td>B</td>
<td>96.2</td>
</tr>
<tr>
<td>C</td>
<td>84.6</td>
</tr>
<tr>
<td>D</td>
<td>91.0</td>
</tr>
<tr>
<td>GDM-M</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>96.8</td>
</tr>
<tr>
<td>B</td>
<td>95.8</td>
</tr>
<tr>
<td>C</td>
<td>87.3</td>
</tr>
<tr>
<td>D</td>
<td>92.0</td>
</tr>
</tbody>
</table>

A summary of average results, across emotions, is given for each channel and OMC-DNN in Table 4-18. The GIM OMC-DNN outperformed the best individual channels by an average of 4.3%, 4.5%, 10.6% and 8.2% for sets A-D. GDM-F had an average improvement of 2.7%, 3.0%, 8.6% and 2.7% for each set. Similarly, each OMC-DNN set improved the SC-DNN in GDM-M by 3.4%, 6.4%, 7.7% and 2.8%.

Table 4-18 AVERAGE TEST ACCURACY (%), ACROSS 6 BINARY PAIRS, OF THE THREE INDIVIDUAL CHANNELS AND THE OVERALL 3-CHANNEL OMC-DNN FOR FEATURE SETS A-D WITH GIM AND GDM

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Individual Channels Accuracy (%)</th>
<th>OMC-DNN Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Channel 1</td>
<td>Channel 2</td>
</tr>
<tr>
<td>GIM</td>
<td>85.2</td>
<td>79.1</td>
</tr>
<tr>
<td>B</td>
<td>81.7</td>
<td>77.8</td>
</tr>
<tr>
<td>C</td>
<td>70.6</td>
<td>70.0</td>
</tr>
<tr>
<td>D</td>
<td>75.3</td>
<td>76.6</td>
</tr>
<tr>
<td>GDM-F</td>
<td>89.2</td>
<td>82.2</td>
</tr>
<tr>
<td>B</td>
<td>84.2</td>
<td>80.3</td>
</tr>
<tr>
<td>C</td>
<td>76.3</td>
<td>75.4</td>
</tr>
<tr>
<td>D</td>
<td>84.0</td>
<td>78.7</td>
</tr>
<tr>
<td>GDM-M</td>
<td>89.5</td>
<td>83.7</td>
</tr>
<tr>
<td>B</td>
<td>80.9</td>
<td>81.8</td>
</tr>
<tr>
<td>C</td>
<td>78.9</td>
<td>75.4</td>
</tr>
<tr>
<td>D</td>
<td>85.0</td>
<td>80.0</td>
</tr>
</tbody>
</table>
Each 3-channel feature set was used in GIM multiclass AER with channel weights given in Table 4-19. The results indicate set A (S-MFCC), set B (G-MFCC) and set D (G-TEO) are significantly weighted in the static coefficient (channel 1) so derivatives (channel 2 and 3) are not that important. Set C the channel 3 (double derivative) is effectively useless and the two remaining channel equally important.

Table 4-19 AVERAGE WEIGHT ALLOCATION, FOR FEATURE SETS A, B, C AND D (STATIC, DERIVATIVE, DOUBLE DERIVATIVE FEATURE CHANNELS), OF THE 3-CHEannel OMC-DNN IN GIM MULTICLASS (7-CLASS) AER

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Average Weight Allocation of independent feature channels</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Channel 1</td>
</tr>
<tr>
<td></td>
<td>Static</td>
</tr>
<tr>
<td>A</td>
<td>S-MFCC</td>
</tr>
<tr>
<td>B</td>
<td>G-MFCC</td>
</tr>
<tr>
<td>C</td>
<td>S-TEO</td>
</tr>
<tr>
<td>D</td>
<td>G-TEO</td>
</tr>
</tbody>
</table>

The accuracy of the 3-channel OMC-DNN for multiclass AER is given by Table 4-20, for sets A, B, C and D. In set A the final result of 58% was the same as the best individual feature channel. This channel (MFCC) was weighted much higher than the other channels so this could explain the result. There was only a small improvement in OMC-DNN compared to the best SC-DNN is 0%, 2%, 1%, and 1% for sets A to D.

Table 4-20 AVERAGE ACCURACY (%) FOR MULTICLASS AER OF INDIVIDUAL SC-DNN AND THE ENTIRE OMC-DNN USING A 3-CHEannel SETUP WITH FOUR DIFFERENT FEATURE SETS (A, B, C AND D) IN A GENDER INDEPENDENT MODEL (GIM)

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Individual Channel Accuracy (%)</th>
<th>OMC-DNN Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Channel 1</td>
<td>Channel 2</td>
</tr>
<tr>
<td>A</td>
<td>S-MFCC</td>
<td>58</td>
</tr>
<tr>
<td>B</td>
<td>G-MFCC</td>
<td>46</td>
</tr>
<tr>
<td>C</td>
<td>S-TEO</td>
<td>24</td>
</tr>
<tr>
<td>D</td>
<td>G-TEO</td>
<td>35</td>
</tr>
</tbody>
</table>

Overall the OMC-DNN outperformed the comparable SC-DNN for all gender models and feature sets. The average accuracy of OMC-DNN binary classification reached 92% and 93% for GIM-F and GIM-M with MFCC. The multiclass OMC-DNN led to an accuracy of 58% for 7 emotions in a GIM.
4.12.1.2 Sets E-G as 4 channel OMC-DNN

The following section investigates the OMC-DNN further with four parallel feature channels in three different feature set combinations, given in Table 4-21. The channels included the single channel features (MFCC, G-MFCC, TEO and G-TEO) forming set E with sets F and G using the respective first and second derivatives.

Table 4-21 OUTLINE OF THE THREE DIFFERENT 4-CHANNEL FEATURE SETS (E-G) FOR THE OMC-DNN EMOTION RECOGNITION EXPERIMENTS

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Channel 1</th>
<th>Channel 2</th>
<th>Channel 3</th>
<th>Channel 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>S-TEO</td>
<td>S-MFCC</td>
<td>G-TEO</td>
<td>G-MFCC</td>
</tr>
<tr>
<td>F</td>
<td>Δ S-TEO</td>
<td>Δ S-MFCC</td>
<td>ΔG-TEO</td>
<td>ΔG-MFCC</td>
</tr>
<tr>
<td>G</td>
<td>ΔΔ S-TEO</td>
<td>ΔΔ S-MFCC</td>
<td>ΔΔG-TEO</td>
<td>ΔΔG-MFCC</td>
</tr>
</tbody>
</table>

The weights given in Table 4-22, averaged for six binary models indicate the highest to lowest ranked features are MFCC, G-MFCC, G-TEO and TEO and follows the trend of SC-DNN performance. The weights of each channel in set E show a large average difference, set F the differences between weights are reduced and more so in set G with almost equal channel weights. This indicates the static features have a wider variation in importance; where as the derivative features are equivalent.

Table 4-22 AVERAGE WEIGHT ALLOCATION, OF SIX BINARY AER PAIRS, FOR FEATURE SETS E, F AND G (S-TEO, S-MFCC, G-TEO AND G-MFCC FEATURE CHANNELS) OF THE 4-CHANNEL OMC-DNN IN GIM AND GDM

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Average Weight Allocation of independent channels</th>
<th>Channel 1</th>
<th>Channel 2</th>
<th>Channel 3</th>
<th>Channel 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>S-TEO</td>
<td>S-MFCC</td>
<td>G-TEO</td>
<td>G-MFCC</td>
</tr>
<tr>
<td>GIM</td>
<td>E Static</td>
<td>0.088</td>
<td>0.552</td>
<td>0.169</td>
<td>0.190</td>
</tr>
<tr>
<td></td>
<td>F Δ</td>
<td>0.130</td>
<td>0.306</td>
<td>0.230</td>
<td>0.329</td>
</tr>
<tr>
<td></td>
<td>G ΔΔ</td>
<td>0.157</td>
<td>0.265</td>
<td>0.287</td>
<td>0.285</td>
</tr>
<tr>
<td>GDM-F</td>
<td>E Static</td>
<td>0.123</td>
<td>0.526</td>
<td>0.174</td>
<td>0.176</td>
</tr>
<tr>
<td></td>
<td>F Δ</td>
<td>0.137</td>
<td>0.411</td>
<td>0.191</td>
<td>0.259</td>
</tr>
<tr>
<td></td>
<td>G ΔΔ</td>
<td>0.197</td>
<td>0.247</td>
<td>0.315</td>
<td>0.245</td>
</tr>
<tr>
<td>GDM-M</td>
<td>E Static</td>
<td>0.083</td>
<td>0.658</td>
<td>0.157</td>
<td>0.103</td>
</tr>
<tr>
<td></td>
<td>F Δ</td>
<td>0.135</td>
<td>0.441</td>
<td>0.184</td>
<td>0.240</td>
</tr>
<tr>
<td></td>
<td>G ΔΔ</td>
<td>0.144</td>
<td>0.369</td>
<td>0.178</td>
<td>0.291</td>
</tr>
</tbody>
</table>

The individual OMC-DNN binary AER results for GIM and GDM are shown in Table 4-23 for 4-channel feature sets (E, F, G) with a top result of 98.3% for GDM-M/Anger/Set E. The average difference of GIM/GDM was 1.8%, 4.6% and 7.3% for GDM-F and 4.2%, 7.8% and 7.9% for GDM-M in sets E, F and G.
Table 4-23 AVERAGE ACCURACY (%) FOR THE BINARY AER TESTS USING THE 4-CHANNEL OMC-DNN WITH THREE DIFFERENT FEATURE SETS E-G FOR BOTH GIM AND GDM

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>OMC-DNN Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Anger</td>
</tr>
<tr>
<td>GIM</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>96.5</td>
</tr>
<tr>
<td>F</td>
<td>94.7</td>
</tr>
<tr>
<td>G</td>
<td>91.7</td>
</tr>
<tr>
<td>GDM-F</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>97.2</td>
</tr>
<tr>
<td>F</td>
<td>95.9</td>
</tr>
<tr>
<td>G</td>
<td>94.1</td>
</tr>
<tr>
<td>GDM-M</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>98.3</td>
</tr>
<tr>
<td>F</td>
<td>96.4</td>
</tr>
<tr>
<td>G</td>
<td>94.1</td>
</tr>
</tbody>
</table>

Table 4-24 gives the average accuracy, across six binary emotions, of the individual SC-DNN and OMC-DNN. The results suggest that for all cases, the new OMC-DNN improved the performance compared to the SC-DNN. In GIM the average improvement of OMC-DNN directly compared to the best SC-DNN was 4%, 6% and 7.2% for sets E, F and G. Similarly, sets E, F and G improved on the best SC-DNN by an average of 2%, 3.8% and 10.6% for GDM-F and 1.1%, 4.9%, 6.7% for GDM-M.

Table 4-24 AVERAGE TEST ACCURACY (%), ACROSS 6 BINARY PAIRS, OF EACH INDIVIDUAL CHANNEL AND THE OVERALL 4 CHANNEL OMC-DNN FOR FEATURE SETS E-G WITH GIM AND GDM

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Individual Channels Accuracy (%)</th>
<th>OMC-DNN Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ch1</td>
<td>Ch2</td>
</tr>
<tr>
<td>GIM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>69.5</td>
<td>87.4</td>
</tr>
<tr>
<td>F</td>
<td>69.6</td>
<td>79.2</td>
</tr>
<tr>
<td>G</td>
<td>67.8</td>
<td>75.8</td>
</tr>
<tr>
<td>GDM-F</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>74.9</td>
<td>91.3</td>
</tr>
<tr>
<td>F</td>
<td>75.0</td>
<td>86.0</td>
</tr>
<tr>
<td>G</td>
<td>72.9</td>
<td>79.6</td>
</tr>
<tr>
<td>GDM-M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>78.5</td>
<td>94.4</td>
</tr>
<tr>
<td>F</td>
<td>78.2</td>
<td>88.1</td>
</tr>
<tr>
<td>G</td>
<td>77.0</td>
<td>84.1</td>
</tr>
</tbody>
</table>
4.13 Discussion and Summary

This chapter investigated AER using acoustic features including the newly proposed 2D images represent MFCC and TEO contours from speech (S-MFCC, S-TEO) and glottal waveforms (G-MFCC, G-TEO). The new features were applied in a proposed OMC-DNN and compared to a DNN and benchmark classifiers (SVM, NN, GMM).

A) Discuss performance of glottal and spectral parameters in AER

Glottal waveform derivation (i.e. G-TEO and G-MFCC) improved the performance of TEO but not MFCC for binary and multiclass AER. The results are consistent with past studies that improved recognition rates using glottal waveform features [105][140] with G-TEO [147][238][65] where as G-MFCC tend not to help[367].

This could be due to G-TEO enhanced capabilities of representing non-linearity airflow during each glottal cycle. The purpose of TEO is to capture the non-linear properties, which is produced at the glottis [68]. So removing the vocal tract effect may be helpful in representing these non-linear glottal characteristics.

B) Compare emotions pairings in recognition

Anger was generally the easiest emotion to detect in binary AER and passive emotions (sadness and boredom) were more difficult to separate from neutral. Multiclass AER had misclassification in certain emotion pairings (i.e. anger/joy or sad/neutral/boredom) with misinterpretation due to acoustic similarities [178][356].
C) Summarize performance of benchmark classifiers compared to DNN

The DNN, with 2D features, performed significantly better than standard acoustic feature vectors in benchmark classifiers (NN, GMM and SVM). The NN using 2D features was significantly worse compared to the DNN and the NN with 1D features.

The NN struggled with 2D features, indicating it cannot learn abstract concepts, where as the DNN was capable of learning complex connections within 2D images even without unsupervised pre-training (DNN-NP). DNN have been shown to learn representations with several levels of abstraction [173].

D) Discuss the relation of gender dependence in AER (GDMs/GIM)

As expected the investigation into gender dependence found that the gender dependent models (GDMs) increase the accuracy of AER for all experimental setups. No trend was in relation to a gender that has better emotion recognition. There were different results depending on the feature, emotion and classification combinations.

E) Compare OMC-DNN to single channel DNN

Using the OMC-DNN with multiple feature channels showed improvement for all of the feature sets, gender models and emotions. This is shown by the compilation of all of the OMC-DNN feature set results Table 4-25. This compares the average accuracy of the binary AER using the OMC-DNN and the top individual channel.

Table 4-25 SUMMARY OF THE BINARY OMC-DNN RESULTS, FOR EVERY FEATURE SET (A, B, C, D, E, F), GIVING THE AVERAGE TEST ACCURACIES AND THE CORRESPONDING TOP INDIVIDUAL FEATURE CHANNEL OF THE SC-DNNs.

<table>
<thead>
<tr>
<th>OMC-DNN Feature sets Average Accuracy (%)</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>GIM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top Channel</td>
<td>85.2</td>
<td>81.7</td>
<td>70.6</td>
<td>76.6</td>
<td>87.4</td>
<td>79.3</td>
<td>75.8</td>
</tr>
<tr>
<td>OMC-DNN</td>
<td>89.6</td>
<td>86.2</td>
<td>81.2</td>
<td>84.8</td>
<td>91.3</td>
<td>85.2</td>
<td>83.0</td>
</tr>
<tr>
<td>GDM-F</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top Channel</td>
<td>89.2</td>
<td>84.2</td>
<td>76.3</td>
<td>83.9</td>
<td>91.3</td>
<td>85.9</td>
<td>79.6</td>
</tr>
<tr>
<td>OMC-DNN</td>
<td>91.9</td>
<td>87.1</td>
<td>84.9</td>
<td>86.7</td>
<td>93.2</td>
<td>89.8</td>
<td>90.2</td>
</tr>
<tr>
<td>GDM-M</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top Channel</td>
<td>89.5</td>
<td>81.8</td>
<td>78.9</td>
<td>85.0</td>
<td>94.4</td>
<td>88.1</td>
<td>84.1</td>
</tr>
<tr>
<td>OMC-DNN</td>
<td>92.9</td>
<td>88.2</td>
<td>86.6</td>
<td>87.8</td>
<td>95.5</td>
<td>93.0</td>
<td>90.8</td>
</tr>
</tbody>
</table>
The best individual channel on average was the first channel for sets A-D (MFCC, G-MFCC, TEO, G-TEO) and the second channel for sets E-G (MFCC, ΔMFCC, ΔΔMFCC). Indicating the importance of the static (not derivatives) and the MFCC (instead of TEO, G-TEO, G-MFCC).

The 3-channel OMC-DNN outperforms the top performing SC-DNN in feature set A by 4%, 3% and 3% for GIM, GDM-F and GDM-M. Similar set B was improved by 5%, 3% and 6%, set C by 11% 9% and 8% and for set D by 8%, 3% and 3%. The 4-channel OMC-DNN outperforms the best SC-DNN for GIM, GDM-F and GDM-M in set E by 4%, 2%, and 1%, set F by 6%, 4% and 5% and set G by 7% 11% and 7%.

The channels weights did not directly correlate to the independent SC-DNN accuracy. Such that a feature channel weighted significantly higher would not necessarily be reflected in channel accuracy. This would suggest channel weights would be affected by the attributes of all parallel channels in the overall system.

F) **Comparison feature sets and channel size of OMC-DNN**

Based on the Table 4-25 summary it was shown that the 4-channel OMC-DNN outperformed the 3-channel system with the best result in GIM (Set E: 91.3%), GDM-M (Set E: 95.5%) and GDM-F (Set E: 93.2%). The average binary accuracies of the OMC-DNN are summarized in Table 4-26 comparing the worst and best feature sets for 3-channel (sets A, B, C, D) and 4-channel (sets E, F, G). This indicates the 4-channel system has a higher accuracy range, compared to 3-channels, at the expense of the increased computational load.

<table>
<thead>
<tr>
<th></th>
<th>MIN</th>
<th>MAX</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3-channel</td>
<td>4-channel</td>
</tr>
<tr>
<td>GIM</td>
<td>81</td>
<td>83</td>
</tr>
<tr>
<td>GDM-F</td>
<td>85</td>
<td>90</td>
</tr>
<tr>
<td>GDM-M</td>
<td>87</td>
<td>91</td>
</tr>
</tbody>
</table>

Table 4-26 **Summary of Binary OMC-DNN Results with the Minimum and Maximum Accuracy of Each Type of Feature Sets Including 3-Channel (A-D) and 4-Channel (E-G)**
G) Best feature classifier setup

The OMC-DNN was the best overall with feature set E (TEO, MFCC, G-TEO, G-MFCC) attaining an average accuracy of 91%, for the GIM, 93% for GDM-F and 96% for GDM-M. 7 multiclass AER the best result used feature set A (MFCC, ΔMFCC, ΔΔMFCC) in the OMC-DNN of 58%. The male SC-DNN with the S-MFCC in the multiclass case achieved a maximum of 61% for the GDM-F using S-MFCC.

H) Comparison to past multiclass AER publications

The AER experiments used the popular EMO-DB, which provides a means to compare the results with previously published related literature. Past multiclass AER has been investigated using a variety of features, classifiers, databases and experimental setups given by a summary of related studies supplied in Table 4-27 including state-of-the-art EMO-DB results. Direct comparisons are not possible with differences in dataset, features, classifiers and experimental setups.

Table 4-27 SUMMARY OF RELATED MULTICLASS EMOTION RECOGNITION PUBLICATIONS OUTLINING THE ACOUSTIC FEATURES, CLASSIFICATION SYSTEM, DATABASE AND ACCURACY (%).

<table>
<thead>
<tr>
<th>Publication</th>
<th>Database</th>
<th>Emotion classes</th>
<th>Feature</th>
<th>Classifier</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vlasenko, et al. (2007) [516]</td>
<td>EMO-DB</td>
<td>7 emotions</td>
<td>1407 LLD (76 with feature selection)</td>
<td>GMM-SVM</td>
<td>89.9%</td>
</tr>
<tr>
<td>Schuller, et al. (2006) [517]</td>
<td>EMO-DB</td>
<td>7 emotions</td>
<td>276 LLD</td>
<td>SVM</td>
<td>84.8%</td>
</tr>
<tr>
<td>Huang, et al. (2013) [227]</td>
<td>EMO-DB</td>
<td>7 emotions</td>
<td>Spectrograms</td>
<td>RBM-SVM</td>
<td>71%</td>
</tr>
<tr>
<td>Kim, et al. (2013) [226]</td>
<td>IEMOCAP</td>
<td>4 emotions</td>
<td>Audio+Visual</td>
<td>2-layer DBN</td>
<td>70.5%</td>
</tr>
<tr>
<td>Georgogiannis, et al. (2012) [189]</td>
<td>EMO-DB</td>
<td>7 emotions</td>
<td>MFCC, ΔMFCC, ΔΔMFCC</td>
<td>GMM</td>
<td>63%</td>
</tr>
<tr>
<td>He, et al. (2009) [235]</td>
<td>ORI-DB</td>
<td>5 emotions</td>
<td>TEO</td>
<td>GMM</td>
<td>62%</td>
</tr>
<tr>
<td>Le, et al. (2013) [225]</td>
<td>FAU Aibo</td>
<td>5 emotions</td>
<td>MFCC, ΔMFCC, ΔΔMFCC</td>
<td>DBN-HMM</td>
<td>46.4%</td>
</tr>
</tbody>
</table>
The results presented in this thesis, with a top multiclass accuracy of 61%, was much lower than state-of-the-art EMO-DB publications that have attained close to 97% accuracy of seven emotions [517]. Although, it has to be noted that the studies by Vlasenko et al. and Schuller et al. used very large feature sets compared to the experiments presented in this work.

Other studies that are more comparable to the approach taken in this thesis, using smaller feature sets and deep learning approaches, have closer results with RBM-SVM or DBNs have utilized images of spectrograms or multimodal with facial images between 70%-74% [226][227].

Studies that relate to the feature approach taken in this thesis, using MFCC or TEO acoustic parameters, has shown comparable results between 46%-63% depending on the classification setup [225][356][235][189].

In particular, a deep learning study by Le and Provost used a DBN-HMM with MFCC and derivatives and achieved a top result of 46.4% for 5-class AER [225]. This could be considered comparable to the OMC-DNN using MFCC, ΔMFCC, ΔΔMFCC feature set presented in this thesis, which performed better with 58% in 7-class AER.
Chapter Five:

CONVERSATION MODELING SYSTEM: INFLUENCE MODEL EXPERIMENTAL RESULTS WITH MODELING/CLASSIFICATION OF DEPRESSION

5.1 Preview
This chapter explores the capacity of using the Influence model (IM) as a conversation modeling system (CMS), capable of capturing emotional interactions. Additionally, the IM parameters combined with machine learning are explored towards a depression classification system.

The chapter starts, in Section 5.3, with an in depth explanation of the IM from theoretical studies and applications. Then new Emotional Influence Models (EIM) and the extensions of dynamic (DEIM) and higher order (HOEIM) implementations are introduced and explained.

The resulting models generated features, Section 5.4, for quantitative results to describe qualitative emotional influence patterns in parent-adolescent conversations and towards classification described in Section 5.4.2. The experimental setup and evaluation methods are outlined in Sections 5.5 and 5.6. Results are supplied and discussed in Sections 5.7-5.9 and concluded in Section 5.10.
5.2 Oregon Research Institute (ORI) annotations corpus

This research was facilitated by ORI-DB, outlined in Section 3.3, using the parent-adolescent dyadic conversation subset (29 depressed and 34 non-depressed) defined in Section 3.3.5 with the related LIFE annotation codes [23].

The LIFE coding system, as explained in Section 3.3.4, are second-by-second behavioral codes that relate to the observed affect designed to capture the affective interpersonal behavior based on 10 codes: contempt, anger, anxious, dysphoric, pleasant, neutral, happy, caring, whine and belligerence.

The LIFE affect codes are not well balanced and due to the finite length of conversational data some affect codes and transitions occur rarely or never. Furthermore the large number of states is problematic leasing to high complexity and computational expense. The solution was to create meta-states of grouped emotions, but there is no unified system to group discrete emotions.

Many different emotional categories have been proposed and reviewed by Cowie et al. in [233]. A common approach is using 3-dimensions of valance, arousal and control (dominance) [231][232] or 2D disregarding the control variable that is only a narrow range [243][205].

Valence represents discrete emotions on a continuous scale from positive to negative and arousal is the intensity [11]. Kuppens et al. used the idea of positive and negative emotional constructs [88][87]. Furthermore, AER studies have justifiably and successfully grouped positive and negative emotions [407].

Emotions such as anger, joy, and fear and to lesser extent, surprise has positive activation and similar acoustic characteristics (higher $F0$ and wider $F0$ range). Disgust, sadness and boredom have negative activation with comparable characteristics (lower $F0$ and narrower $F0$ range) [356].
The similarity of acoustic features for particular emotions indicates they can be mistaken. It is recommended that grouping emotions with similar characteristics may improve performance [400].

Silence conveys important information pertaining to speech fluency with time intervals of speech and silence of interest in conversational behavior and can invoke negative emotions [142]. Zeligs et al. stated silence can reflect many emotions and can be a sign of contentment, mutual understanding, compassion, emptiness or lack of affect [141] and is a strong correlation of depression [98]. Turn taking is a popular choice in IM using speech and silence as states [80][81][85][84][83][86]. For these reasons silence should be included for analysis.

The final four labels used in this thesis, in the EIM and affect forecasting experiments, were manually generated from the supplied LIFE annotations outlined in Section 3.3.4, and defined as follows:

1) Positive (+): pleasant, happy or caring

2) Negative (-): contempt, anger, belligerence, anxious, dysphoric, or whine

3) Neutral (n): neutral emotion

4) Silence (s): no speech activity;

5.2.1 ORI-DB Annotation Corpus– Statistical Analysis

The ORI-DB annotation system used in the experiments is based on second-by-second states belonging to the set of constructs including positive, negative, neutral and silence. A preliminary analysis of the annotation sequences of each participant in the parent-adolescent conversations was conducted to examine basic statistical information related to the frequency and duration of each construct state.
The duration of the four construct states (positive, negative, neutral and silence) was analyzed based on the average percent of time a speaker spent in each construct. The results are supplied in Figure 5-1 averaged over the three interactions (EPI, FCI, PSI) for each gender and depressed status of adolescents and their parents.

![Average proportion of time spent in each construct](image)

**Figure 5-1** Average proportion (%) of time adolescents and parents spent in each construct state comparing conversations with depressed and non-depressed adolescents averaged for both genders and all interactions (EPI, FCI and PSI).

The frequency of the four construct states was defined as the average time (seconds) of the consecutive sequences of each construct state. The results are given in Figure 5-2 averaged across each speaker and interaction for each gender comparing the depressed and non-depressed adolescents and their parents.

![Average duration of consecutive constructs](image)

**Figure 5-2** Total average duration (seconds) consecutively spent in each construct state comparing conversations with depressed and non-depressed adolescents averaged over both genders and all interactions (EPI, FCI and PSI).
The proportion results indicated silence was the most frequent state, more so with adolescents, with turn-taking approximately half of the conversation. Neutral was the most common speech construct, especially in the case of parents. Where as, the adolescents had more positive and negative states than parents. Also depressed adolescents were more frequently negative and positive compared to non-depressed

It was observed that the average duration of consecutive constructs was 3-6 seconds depending on the construct. This agrees with current literature related to emotion duration [507][508][509]. The results indicated silence was the longest sequence in adolescents and neutral in parents. Depressed adolescents had shorter neutral and longer negative and positive sequences compared to non-depressed.

Parents of depressed adolescents had an increase/decrease in frequency and duration of negative/positive states compared to parents of non-depressed. The elevated levels of negative interactions and absence of positive interactions is a negative reciprocal process associated with reinforcing adolescent depression [137][135][144][24][136].

The observed increased frequency and duration of negative states in depressed adolescents agrees with current studies that recognize increases in anger and sadness in depression [445][449] and difficulties regulating negative emotions [137][24]. The increased frequency and duration of both positive and negative constructs in depressed adolescents is supported by evidence that expressed emotions are more evident during depression [310][311] with resistance to emotional change [87][88].

The observations presented, based on construct duration and frequency, suggest valuable knowledge in conversations related to depression. These basic statistics provided no information regarding construct transitions that could be analyzed in a complex statistical model.
5.3 Influence Model background knowledge

As previously mentioned in the literature review, Section 2.4.3, there are many appropriate models for analyzing interactions. A HMM could model each participant in a conversation but is complex and computationally expensive. Problematically, the total number of states increases exponentially with the number of chains, \( M \), such that \( Q^M \) where \( Q \) is the number of states per chain and \( Q^{2M} \) transition matrices parameters.

The solution is a coupled HMM, able to determine a chain’s state influenced by all chains previous states [283][284], shown by Figure 5-3[514]. This model reduces dimensionality, estimating \( P(S_t^1|S_{t-1}^1,\ldots,S_{t-1}^M) \), requiring \( Q^M \times Q \) transition tables per chain and a total of \( MQ^{M+1} \) transition parameters per chain and is still large.

![Generalized coupled HMM for \( M \) chains for interaction between \( t-1 \) and \( t \) with states, \( S \) and hidden states, \( x \).](image)

Figure 5-3 Generalized coupled HMM for \( M \) chains for interaction between \( t-1 \) and \( t \) with states, \( S \) and hidden states, \( x \).
5.3.1 Introduction to the Influence Model

Asavathiratham introduced the Influence Model as a tractable mathematical representation of random, dynamical interaction on networks [26]. It is a generative model for describing interactions between Markov chains with a parameterization in terms of the “influence” each chain has on the others.

The states in the IM are all considered observable where the graphical representation of the IM is identical to a generalized M-chain coupled HMM expect for a simplification by approximating the joint-conditional probabilities as follows:

\[
P(S_t^i|S_{t-1}^1, ..., S_{t-1}^M) = \sum \theta_{ij} P(S_t^i|S_{t-1}^j) \quad (5.1)
\]

The IM characterizes a sequence of states, \( \{S_t^i\} \), where the distribution for a given chain’s, \( i \), next state is given by equation (5.1) as a convex combination of pairwise conditional probabilities, \( P(S_t^i|S_{t-1}^j) \), weighted by constant parameters, \( \theta_{ij} \), for \( 1 \leq i, j \leq M \) and \( M \) chains [26]. Such that \( P(S_t^i|S_{t-1}^j) \) denotes the conditional probability that chain \( i \) is in state \( S \) at time \( t \) given chain \( j \) was in a given state at \( t-1 \). The weights portray the self-influence each chain has on itself \( \theta_{ii} \) and cross-influence from other chains \( \theta_{ij} \) and are normalized such that \( \sum_j \theta = 1 \) with \( \theta > 0 \).

This reduces the parameters per chain to \( M Q \times Q \) tables and \( M \theta_{ij} \), far less than HMM. An advantage is a relatively small set of parameters that can be easily interoperated for psychological analysis. This is at the cost of losing modeling power, while the fully connected coupled HMM allows for explicit modeling of joint events.

The weight parameters, \( \theta \), linking each chain describe speaker interaction in terms of how much they influence each other. The constant parameters, known as Influence Coefficients (IC), describe how much influence each neighbor has on neighboring nodes and the transition probabilities describe how a chain is influenced.
5.3.2 Learning the influence Model (Optimization Algorithm)

The parameters, $\theta_{ij}$, and $P(S_i^j|S_{i-1}^j)$, of the IM for a given chain, $i$, of observations, $\{s_t^i\}$ is found using an optimization algorithm. The IM is fit to the data by maximizing the likelihood over the free parameters using the EM algorithm used by the majority IM applications [84][85][86][81][83][80].

The EM algorithm is an iterative method for finding the maximum likelihood or maximum posterior estimates of parameters. The EM procedure alternates between the expectation step (E) and maximization step (M). The E-step creates a function for the expectation of the log-likelihood using current parameter estimates and the M-step calculates new parameters maximizing the expected log-likelihood from the E-step and used to determine the distribution of latent variables in the next E-step.

The likelihood function is given by equation (5.2) and to estimate the parameters of this model the E-step requires calculation of $P(S|X, \theta)$ using an approximate inference scheme (i.e. Variable Bayesian method).
\[ P(S, X) = \left( \prod_{i=1}^{M} P(S_0^i)P(x_0^i|S_0^i) \right) \left( \prod_{i=1}^{M} \prod_{t=1}^{T} P(x_t^i|S_t^i) \right) \sum_{i} \theta_{ij} P(S_t^i|S_{t-1}^j) \] (5.2)

The M-step in this model requires maximizing the lower bound obtained in the E-step. The update equation for parameters, \( \theta_{ij}^{new} \), as a mixture weights for \( M \) conditional probability matrices similar to a mixture of Gaussians.

\[ \theta_{ij}^{new} = \frac{\sum_{t} \sum_{k} \sum_{l} P(c_t^i = j, S_t^i = k, S_{t-1}^j = l | X)}{\sum_{t} \sum_{k} \sum_{l} P(S_t^i = k, S_{t-1}^j = l | X)} \] (5.3)

\( c_t^i = j \) event means at time \( t \) chain \( i \) influence by \( j \)

\( S_t^i = k \) event means chain \( i \) was in state \( k \) at time \( t \)

A problem is the difficulty involved in solving the inference required at the E-step and it is reasonable to simplify by only allowing observed states in each chain, shown in Figure 5-3. The \( \theta_{ij} \) in the observed IM are estimated with a simplified likelihood function, which prevents over fitting found in the full model [84]. The likelihood function for the observed IM is given by equation (5.4). It is easier to use the logarithm, equation (5.5), to replace the products with summation.

\[ P(S|\theta_{ij}) = \prod_{i=1}^{M} P(S_0^i) \prod_{i=1}^{M} \prod_{t=1}^{T} \sum_{j} \theta_{ij} P(S_t^i|S_{t-1}^j) \] (5.4)

\[ P(S|\theta_{ij}) = \prod_{i=1}^{M} P(S_t^i) \sum_{i=1}^{M} \sum_{t=1}^{T} \log \sum_{j=1}^{M} \theta_{ij} P(S_t^i|S_{t-1}^j) \] (5.5)

The aim is to maximize the likelihood of a sequence by optimizing the parameters in over \( \theta_{ij} \) with irrelevant terms removed, given by (5.6) equation, and simplifying the log-likelihood function for a given chain \( i \) defined by equation (5.7).
\[
P(S|\theta_{ij}) = \sum_{i=1}^{M} \sum_{t=1}^{T} \log \sum_{j=1}^{M} \theta_{ij} P(S_i^t|S_{t-1}^j) \tag{5.6}
\]

\[
P(S|\theta_{ij}) = \sum_{t=1}^{T} \log \sum_{j=1}^{M} \theta_{ij} P(S_i^t|S_{t-1}^j) \tag{5.7}
\]

The objective is to find, \(\theta_{ij}\) and \(P(S_i^t|S_{t-1}^j)\), that maximize the concave log-likelihood function, given by equation (5.8). For a given speaker \(i\) the parameters, \(\theta_{ij}\), can be determined with the EM algorithm [294].

\[
\theta_{ij}^* = \arg\max_{\theta_{ij}} \left[ \sum_{t} \log \sum_{j} \theta_{ij} P(S_i^t|S_{t-1}^j) \right] \tag{5.8}
\]

The EM using the log-likelihood is beneficial over minimizing the MSE, which heavily weights outliers [179]. Although other methods have been compared including gradient decent, multivariate linear regression (MVR), gradient decent (GD), simulated annealing (SA), Quasi-Newton (QN) simulated annealing with genetic algorithm (SA/GA). Depending on the nature of the optimization technique, two types of constraints have been used: Type 1 constraint limiting the coefficient’s values to the range \((0, 1)\) and Type 2 constraint normalizing to sum to 1 and are defined in Table 5-1 along with the objective function (MSE or Log-Likelihood).

<table>
<thead>
<tr>
<th>Method</th>
<th>Objective function</th>
<th>Constraints</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Type 1</td>
</tr>
<tr>
<td>EM</td>
<td>LL</td>
<td>-</td>
</tr>
<tr>
<td>MVR</td>
<td>MSE</td>
<td>-</td>
</tr>
<tr>
<td>GD</td>
<td>MSE</td>
<td>-</td>
</tr>
<tr>
<td>QN</td>
<td>MSE</td>
<td>X</td>
</tr>
<tr>
<td>SA/GA</td>
<td>MSE</td>
<td>X</td>
</tr>
<tr>
<td>SA</td>
<td>MSE</td>
<td>X</td>
</tr>
</tbody>
</table>
5.3.3 *Introducing emotional interactions to Influence Models (EIM)*

The IM represents parent-adolescent conversations as interacting coupled Markov chains, shown by Figure 5-5, linking dependencies between chains. It is assumed the current state of a speaker at time \( t \) depends only on their previous state and the other speakers' state at \( t-1 \) with no other external influences.

In this application the IM was renamed EIM due to the emotional states of the chains. The influence coefficients \((\text{IC})\) describe how much influence and the emotional transitions describe how they are influenced emotionally. It should be noted the word “influence” in this thesis refers to IC and could differ from the psychological definition of this term.

![Figure 5-5 Parent-Adolescent conversation represented by two coupled time chains as a general Influence Model (IM). Each node denotes the state a speaker is at time \( t \). The system models the influence of each speaker on themselves (self-influence \( \theta_{ii} \)) and by others (cross-influence \( \theta_{ij} \)) with time delay \( t-1 \) between each transition.](image)

The Markov chains are defined as two sequences representing states of a parent and adolescent. The EIM is given by the following equations for the adolescent \((i=1)\) and parent \((i=2)\) with time delay of \( n=1 \).

\[
\hat{P}(S^1_t | S^1_{t-1}, S^2_{t-1}) = \theta_{11} P(S^1_t | S^1_{t-1}) + \theta_{12} P(S^1_t | S^2_{t-1})
\]  \hspace{1cm} (5.9)

\[
\hat{P}(S^2_t | S^1_{t-1}, S^2_{t-1}) = \theta_{22} P(S^2_t | S^2_{t-1}) + \theta_{21} P(S^2_t | S^1_{t-1})
\]  \hspace{1cm} (5.10)
The speaker states $S_t^i$ in equations (5.9) and (5.10) were given as annotation labels defined in Section 5.2 with the following set of possible states:

$$S_t^i \in \{\text{Neutral}(N), \text{Positive}(+), \text{Negative}(-), \text{Silence}(s)\}$$

The IC provides quantitative measures to represent intra-speaker influences (self-influence) $\theta_{ij}$ ($i=j$) and inter-speaker influences (cross-influence) $\theta_{ij}$ ($i \neq j$) (influence speaker $j$ on $i$) to express dependencies between and within speakers.

The speakers’ emotional transition influences was given by the probability matrices $P(S_t^i|S_{t-1}^j)$ containing the values of self ($P_{ii}$) and cross ($P_{ij}$) conditional probabilities of described for each dyadic conversation by, $P_{11}, P_{12}, P_{21}, P_{22}$.

The IM/EIM assume that the current state of a speaker at time $t$ depends only on immediate previous state, $t-1$, and other state transitions are not taken into account. The time delay between nodes in a Markov chain is an arbitrary lag length and should considered for the particular application and the choice imposes challenges.

Dynamic timing information is important in psychological analysis to provide information characterizing a person’s mental state and ability to react to external influences [503]. It has been found that longer delays between subsequent states are more important for emotional analysis and psychological information [87][88].
5.3.4 Introducing a Dynamic Emotional Influence Model (DEIM)

To compensate for the deficiency in the time delay limitation of the IM/EIM it was extended into a Dynamic Emotional Influence Model (DEIM) that generated a trajectory of parameters over a range of time delays shown by Figure 5-6. The new DEIM parameters independently account for transitions at multiple time delays of \( t-n \).

![Figure 5-6 Illustration of Dynamic Emotional Influence Model (DEIM) depicting a dyadic parent-adolescent conversation as an interaction between Markov chains, with nodes denoted by the construct states (positive, negative, neutral silence) with arbitrary time delay \((t-n)\) between transitions](image)

This method observes how the IC changes across different time delays given by equation (5.11). A trajectory of ICs, \( \theta_i(n) \), are given as a function of the time delay, \( n=1,2,...,N \) where \( N \) is the maximum delay of the DEIM. For each delay, \( n \), the optimization algorithm solves finds the best fitting set of the ICs \( \theta_i^{opt}(n) = \{\theta_{ij}^{opt}\} \).

\[
\theta_{ij}(n)\]

DEIM creates a trajectory of IC and probabilities to describe how conversation dynamics evolve over time given by (5.12) and (5.13) for each speaker. The self-influence trajectory, \( \theta_i(n) \), reflects the extent a person, at time \( t \), is influenced by their previous emotion states independently at \( t-1, t-2,...,t-N \). Similarly the cross-influence trajectory \( \theta_{ij}(n) \) describes influence in responsive to others past states.
\[
\hat{P}(S^1_t | S^1_{t-n}, S^2_{t-n}) = \theta_{11}(n)P(S^1_t | S^1_{t-n}) + \theta_{12}(n)P(S^1_t | S^2_{t-n}) \\
\hat{P}(S^2_t | S^1_{t-n}, S^2_{t-n}) = \theta_{22}(n)P(S^2_t | S^2_{t-n}) + \theta_{23}(n)P(S^2_t | S^1_{t-n})
\] (5.12)

This method observes how the IC values change across different time delays but a disadvantage is the DEIM needs to be solved \( n \) times. Moreover each DEIM is only a first order Markov representation. In this approach for a given time delay \( n \), the current state is assumed to be affected only by one previous state \((S_{t-n})\), and the effects of state transitions occur between \( t \) and \( t-n \) are not taken into account.

### 5.3.5 Introducing a Higher Order Emotional Influence Model (HOEIM)

To overcome this issue, the IM/EIM was extended to a full memory Higher Order Emotional Influence Model (HOEIM). This takes into account a longer history of past states so the speaker state is influenced by the previous \( N \) states. Equation (5.14) gives a general form of the HOEIM with \( M \) chains.

\[
\hat{P}(S^i_t | S^1_{t-1}, ..., S^M_{t-1}, S^1_{t-2}, ..., S^M_{t-2}, ..., S^1_{t-N}, ..., S^M_{t-N}) = \sum_{j=1}^{M} \theta_{ij}P(S^j_t | S^j_{t-1}, ..., S^j_{t-N})
\] (5.14)

The entire delay range are represented by delays for \( n = 1, ..., N \) where \( N \) is the order (i.e. \( t-1, t-2, ..., t-N \)). For a given order the HOEIM generates a set of ICs that are constrained such that delays such that \( \sum_j \theta_{ij} = 1 \) and \( \theta_{ij} > 0 \) for \( 1 \leq i,j \leq M \).

\[
\hat{P}(S^i_t | S^i_{t-1}, S^i_{t-2}, ..., S^i_{t-N}) = \theta_{ij}P(S^j_t | S^j_{t-1}, ..., S^j_{t-N}) + \theta_{ij}P(S^j_t | S^j_{t-1}, ..., S^j_{t-N})
\] (5.15)
Figure 5-7 represents an application of the HOEIM for a dyadic conversation so the current state of a speaker at time $t$, depend on his or her own and their counterpart’s previous $N$ states given at times $t-1$, $t-2$, ..., $t-N$.

![Diagram of HOEIM for parent-adolescent conversation](image)

Equations (5.16) and (5.17) describe an $N^{th}$ order HOEIM for each speaker in the dyadic conversation. Techniques such as the Mixed Memory Markov (MMM) model or the N-grams are applied to improve the efficiency higher order estimations. The relatively large number of computations required by this procedure can be outweighed as the HOEIM offers important new insights into conversation analysis.

$$
\hat{P}(S^1_t|S^1_{t-1}, S^2_{t-1}, S^1_{t-2}, S^2_{t-2}, \ldots, S^1_{t-N}, S^2_{t-N}) = 
\theta_{11}P(S^1_t|S^1_{t-1}, S^2_{t-2}, \ldots, S^1_{t-N}, S^2_{t-N}) + \theta_{12}P(S^1_t|S^1_{t-1}, S^2_{t-2}, \ldots, S^1_{t-N}, S^2_{t-N})
$$

(5.16)

$$
\hat{P}(S^2_t|S^1_{t-1}, S^2_{t-1}, S^1_{t-2}, S^2_{t-2}, \ldots, S^1_{t-N}, S^2_{t-N}) = 
\theta_{21}P(S^2_t|S^1_{t-1}, S^2_{t-2}, \ldots, S^1_{t-N}, S^2_{t-N}) + \theta_{22}P(S^2_t|S^1_{t-1}, S^2_{t-2}, \ldots, S^2_{t-N})
$$

(5.17)
5.3.5.1 Higher Order EIM using Mixed Memory Markov Model (HOEIM-MMM)

The Mixed Memory Markov Model (MMM) can be used represent the higher order Markov chain in smaller state space of first order models. The MMM is a weighted linear sum of the first order Markov chains for different delay lengths [29][30][31]. The MMM is described by equation (5.18) for any general conditional probability matrix, $P_{ij}$. To solve the HOEIM each of the conditional probabilities $(i,j = 1,2)$ need to be estimated in the higher order using equation (5.18).

$$P(s^i_t | s^j_{t-1}, s^j_{t-2}, ..., s^j_{t-N_m}) = \sum_{n_m=1}^{N_m} \omega_{ij}(n_m)P(s^i_t | s^j_{t-n_m})$$  \hfill (5.18)

The $N^{th}$ order conditional probability is the weighted sum of the first order conditional probabilities, $P(s^i_t | s^j_{t-n_m})$, as a function of the delay, $t-n_m$. The MMM weight coefficients, $\omega_{ij}(n_m)$, estimate the effect of the past observations over a range of time delays, $n_m = 1, ..., N_m$, and determine the most important time delays. The MMM weights are constrained so $\sum_{n_m=1}^{N_m} \omega_{ij}(n_m) = 1$ and $\omega_{ij}(n_m) > 0$.

Given the first-order representation of the higher order conditional probabilities, the best fitting (optimal) set of the weights $(\omega_{ij}^{opt}(n_m)$ for $n_m = 1, ..., N_m$) for each of the first-order components of the HOEIM was determined using the EM algorithm to maximize the log likelihood function $g(\omega)$ given in equation (5.19), explicitly the objective is to find the estimated, weight values $\omega_{ij}^{*}(n_m)$, that maximize the likelihood function, $g(\omega)$, shown by equation (5.20).

$$g(\omega) = \sum_t \log \sum_{n_m} \omega_{ij}(n_m)P(s^i_t | s^j_{t-n_m})$$  \hfill (5.19)

$$\omega_{ij}^{*}(n_m) = \arg\max_{\omega_{ij}(n_m)} \left[ \sum_t \log \sum_{n_m} \omega_{ij}(n_m)P(s^i_t | s^j_{t-n_m}) \right]$$  \hfill (5.20)
5.3.5.2 Higher Order EIM using n-gram method (HOEIM-ngram)

The MMM weights a series of 1st order conditional probabilities to estimate higher order conditional probabilities. An alternative method, using n-grams, weights all past delays equally.

The n-gram model estimates the higher order conditional probabilities of the HOEIM (i.e. HOEIM-ngram). The $N$ most recent state transitions concatenated are into a single meta-state as a $N$th order Markov model represented as a 1st order Markov model. The conditional probabilities, from equations (5.16) and (5.17) are estimated from the data using relative frequencies of occurrence defined as follows:

$$P(S_t^j|S_{t-n_1}^j, S_{t-n_2}^j, ..., S_{t-n_N}^j) = \frac{C(S_t^j, S_{t-n_1}^j, S_{t-n_2}^j, ..., S_{t-n_N}^j)}{C(S_{t-n_1}^j, S_{t-n_2}^j, ..., S_{t-n_N}^j)}$$ (5.21)

The $C(\cdot)$ terms denote the relative frequencies of speaker states directly from the frequency counts. n-gram probabilities are not generally estimated this way due to problems with n-grams that have very low counts or do not occur at all in analyzed sequences. The estimates incorrectly bias the model because rare transitions can occur more frequently in real distributions (or larger datasets).

To solve this data sparseness problem, data smoothing distributes probability mass towards unseen states or state sequences (n-grams). As the model order increases, the size of probability arrays and possible n-grams increases exponentially, increasing the data sparseness problem.
5.3.5.3 HOEIM Modified Kneser-Ney smoothed ngram (HOEIM-KN-ngram)

Data sparseness has widely been studied and many proposed solutions exist [32][33][34]. Instead of estimating probabilities from data, smoothing methods redistribute probability mass from high occurring n-grams those rare or unseen.

Kneser-Ney (KN) combines multiple smoothing techniques including interpolation and absolute discounting and is widely recognized as the best smoothing method [35]. KN smoothing is an extension of absolute discounting with an efficient lower order (back-off) model. This technique assumes lower order models are significant only with a zero or small count in higher models and are not discounted. KN smoothing iteratively estimates higher order transition by interpolating between raw (unsmoothed) data for n-gram and smoothed probabilities for (n-1)-gram model.

This application uses an improved variation of KN smoothing, introduced by Chen and Goodman, known as modified Kneser-Ney (mKN) and reported to be particularly suitable to support n-gram smoothing [36].

\[
\hat{P}(g|h) = \begin{cases} 
\frac{C(h, g) - D_c}{C(h)} + \gamma(h)\hat{P}_{\text{back}}(g|h) & \text{for } C(h, g) > 0 \\
\gamma(h)\hat{P}_{\text{back}}(g|h) & \text{for } C(h, g) = 0 \text{ and } C(h) > 0 \\
\hat{P}_{\text{back}}(g|h) & \text{for } C(h) = 0 
\end{cases} 
\]  \quad (5.22)

The mKN-n-gram model, given by equation (5.22), estimates the conditional probability \( \hat{P}(g|h) \) of a speaker being in state \( g \) given the previous \( N-1 \) states denoted as \( h \). Where \( \hat{P}(g|h) \) is estimated as the number of times state \( g \) follows the previous \( N-1 \) states \( (h) \). For example the conditional probabilities of \( P(S_t|S_{t-1}) \) are calculated for a bigram (second order N-gram) or \( P(S_t|S_{t-1}, S_{t-2}) \) for a tri-gram (third order n-gram).
mKN smoothing has conditions that define how mass is redistributed. If the context \( h \) never occurs \((C(h) = 0)\), the lower order distribution denoted as the “back off” probability \( \hat{P}_{\text{back}}(g|h) \), is estimated such that the “back off” distribution is a one level lower-order smooth probability. For example, for a trigram model, \( \hat{P}_{\text{back}}(g|h) \) represents the bigram model that has already undergone the smoothing process.

If the context has occurred \((C(h) > 0)\) but the entire n-gram is zero \((C(h,g) = 0)\), the “back off probability” \( \hat{P}_{\text{back}}(g|h) \) is weighted by a constant parameter \( \gamma(h) \) to ensure the conditional distribution sums to 1. When the n-gram has been observed (i.e. \( C(h,g) > 0 \)), the interpolated model uses the lower order counts, however in this case the estimate includes a discounting term \( D_c \).

\[
D = \begin{cases} 
1 - 2Y \left( \frac{n_2}{n_1} \right), & \text{if } c = 1 \\
2 - 3Y \left( \frac{n_3}{n_2} \right), & \text{if } c = 2 \\
3 - 4Y \left( \frac{n_4}{n_3} \right), & \text{if } c \geq 3 
\end{cases} \quad (5.23)
\]

\[
Y = \frac{n_1}{n_1 + 2 \times n_2} \quad (5.24)
\]

KN smoothing has a constant discount, although for modified KN smoothing it depends on the n-gram count. Three discount constants for one-count, two-count and three or more counts of the ngram are required as defined by equation (5.23). Where \( Y \) is calculated using equation (5.24) and \( n1 \) and \( n2 \) are the total number of n-grams with one and two count respectively and so on for \( n3 \) and \( n4 \). More details on the N-Gram approach and the Modified KN smoothing can be found in [36].
5.4 Depression Detection using Influence Models

Figure 5-8 illustrates the general flowchart of the depression detection system, used for experiments, including two phases. The first phase was to train features, with EIM parameters \( \{\phi, P\} \), from known samples to learn statistical models of each class. The second phase used the models to classify/test unknown samples to measure performance using accuracy, sensitivity and specificity.

5.4.1 Feature Extraction from the Influence Models

The IM computed a set of parameters for both the parent and adolescent participants, which can be considered a conversational signature of the interaction \( \{\Phi, P\} \) where; \( \Phi \) are the set of four IC \( \{\theta_{11}, \theta_{12}, \theta_{21}, \theta_{22}\} \) and P defines the transition (conditional) matrices \( \{P_{11}, P_{12}, P_{21}, P_{22}\} \). The number of ICs is set for each chain (speaker) as the number of chains (speakers) in the system, which in this case was two (dyadic conversation) per chain. The number of conditional probability parameters in each probability matrix is related to the number of speakers and the number of states as well as the order for the HOEIM cases. In general for each chain the number of ICs, transition probabilities and the total number of parameters are described by Table 5-2.
Analysis of the features can be quantitative using various approaches (t-test, ANOVA, MANOVA, mRMR) to find significant differences between depressed and control groups. From the results qualitative psychological assessments can be made to interoperate emotional influence and dynamics exhibited by depressed adolescents. Additionally, analyzing the features in comparison to the topics (EPI, PSI and FCI) and the orders of the different EIMs can show details about depression.

Table 5-2 Number of Influence Model Parameters (Influence Coefficients (IC) and Transition Probabilities (P)) in each Approach, Where M is the Number of Speakers, Q is the Number of States and N is the Order

<table>
<thead>
<tr>
<th></th>
<th>DEIM</th>
<th>HOEIM-MMM</th>
<th>HOEIM-NGRAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>IC Per Chain</td>
<td>$M$</td>
<td>$M$</td>
<td>$M$</td>
</tr>
<tr>
<td>P Per Chain</td>
<td>$MQ^2$</td>
<td>$N(MQ^2)$</td>
<td>$MQ^{N+1}$</td>
</tr>
<tr>
<td>Per Chain Total</td>
<td>$M+MQ^2$</td>
<td>$M+N(MQ^2)$</td>
<td>$M+MQ^{N+1}$</td>
</tr>
<tr>
<td>Total ${\Phi, P}$</td>
<td>$M(M+MQ^2)$</td>
<td>$M(M+N(MQ^2))$</td>
<td>$M(M+MQ^{N+1})$</td>
</tr>
</tbody>
</table>

5.4.2 Modeling and Classification Setup

The features extracted from the DEIM and HOEIM, $\{\Phi, P\}$, can be used to determine risk factors of depression and discriminate between depressed or control subjects. The features are used to train within machine learning so unknown subjects can be classified using the learned models. The classification system objectively determines the probability of depression in adolescents based on the emotional influence patterns.

Three automatic classification generated models of two classes Depressed and control using: Support Vector Machine (SVM) and Neural Network (NN) that was further explored using optimized number of hidden nodes (optimized NN).

5.4.2.1 Support Vector Machine (SVM)

The SVM was trained using a linear kernel with sequential minimization optimization (SMO). After, preliminary experiments showed minimal difference between leave one out CV and k-fold CV for efficiency the experimental setup used 5-fold CV with 80% training and 20% test sets with different mutually exclusive subsets.
5.4.2.2 Neural Network (NN)

The Neural Network (NN) used to classify depression from EIM features follows the theory explained in Section 4.6.3, designed with a single 25 neuron hidden layer and a tan-sigmoid activation function. The output activation function is tan-sigmoid with two outputs, representing each class. The NN parameters were learned using the Levenberg-Marquardt optimization algorithm evaluated using MSE. The data was segmented into 70%/15%/15% for train, validation and test sets and implemented early stopping CV to ensure training ends when generalization stops improving and averaged over k-fold CV test performances.

5.4.2.3 Optimized Neural Network (o-NN)

The design of the input and output layers of the NN is generally a simple task, the number of input neurons is related to the dimension of the features and the output based on the desired outcome of the network such as the number of classes. However, the network topology of hidden layers, number of layers and number of hidden nodes, are not as trivial. Generally one hidden layer is sufficient for most problems [344][316] but the number of hidden nodes greatly affects performance [406][408].

The generalization of NNs is based on over fitting, which is related to the capacity of the network, determined by the number of weights [150] and variance of the training data [151]. Appropriate network parameters theoretically could be determined but the practical capacity is less than in theory[152][153]. If there are not enough hidden units, there could be high training error and high generalization error due to under fitting and high statistical bias. Too many hidden neurons can degrade generalization due to over fitting and high variance [149].
Many studies and theories suggest formulae or ‘rules of thumb’ to determine the number of hidden nodes [312][313][314] but basic rules may not be optimal [502]. It’s not feasible to determine the architecture for the hidden layers with simple rules and the complexity of the architecture depends on multiple factors.

Studies have proposed various methods to optimize the hidden layers of NN [340]. A popular choice is pruning, which iteratively updates connections based on the weight, by removing unnecessary connection. A common method to determine the number of hidden neurons is with cross-validation and early-stopping [317][318].

Figure 5-9 Procedure to optimize number of hidden neurons in the NN such that the aim to find $H_{opt}$ that results in the highest average CV accuracy, $A^{avg}$.

In this application a range of hidden nodes was explored from 10 to 50 at 5 neuron intervals, illustrated by the procedure given in Figure 5-9. For each setup the data was segmented into 70% training, 15% validation and 15% test sets with early-stopping CV. This process was repeated 5 times in k-fold CV and the final test results were averaged. By applying this method the final architecture of the network, in terms of hidden layer size, was optimized to improve the classification performance.
5.5 Experimental Setup

The experiments in this chapter use the database detailed in Chapter Three and the corpus summarized in Section 3.3.5. The overall aim was an overall system that determined if an adolescent in a conversation is depressed or non-depressed from conversational parameters using the Influence Model. The process of extracting EIM parameters and methodology of classifiers are outlined in Sections 5.3 to 5.4.2. The experiments were conducted on three main experimental categories as follows:

- Influence model experiments
- Depression based analysis of EIM parameters
- Depression classification experiments with EIM

The EIM experiments were intended to compare parameter optimization approaches and research multiple EIM variations including higher order dependence. The experiments are briefly explained below with more detail in the results section.

- **EXPprelim**: Higher order dependence in the influence model initially examined using auto and cross correlation of the annotated LIFE constructs.

- **EXP1**: Compared the parameters derived for the EIM using different optimizations methods based on CPU time and MSE.

- **EXP2**: Experimented with effect of time delay of conversations with a new Dynamic Emotional Influence Model (DEIM) using the best optimizer found in **EXP1**. Compare fit of DEIM delay lengths based on the Log-Likelihood.

- **EXP3**: Investigated higher order benefits of the EIM and examined different approaches of generating the new HOEIM comparing MMM and n-gram.
Experiments conducted on the EIM features intended to supply validation of psychological ideas with the quantitative parameters related to qualitative emotional influences, with the experiments briefly outlined as follows:

- **EXP4** Observed statistically significant EIM probabilities and make qualitative responses of psychological characteristics that differ in depressed.

- **EXP5**: Repeated **EXP4** for the Influence Coefficients from each of the EIM for the DEIM and HOEIM variations.

The next experimental category involves classification of depression using the EIM parameters, based on previous experiments, and machine learning models.

- **EXP6**: Used all features from DEIM derived in **EXP2** to train a SVM and use the models to test classification of depression comparing the effect of delay.

- **EXP7**: Compared the performance of the SVM classifiers for depression detection with an NN using all of the features from the DEIM.

- **EXP8**: Considering the NN performed well an optimum version to find ideal hidden nodes was used to improve the accuracy with the best EIM.

- **EXP9**: Investigated higher order EIM parameters, from **EXP3**, for SVM depression classification and compared KN-ngram and MMM approaches.

- **EXP01**: Investigated a feature selection strategy implemented with a 2-stage mRMR approach on the large HOEIM feature set. Specifically the KN-ngram HOEIM parameters were examined continuing on from **EXP9** that showed improved performance compared to he MMM implemented HOEIM.
5.6 Evaluation Methods

The EIM was evaluated on CPU time and the LL, outlined in Section 5.3, to determine the best optimization method and EIM (i.e. DEIM/HOEIM and delay/order). Depression classification was evaluated on accuracy, sensitivity and specificity given by equations (5.25), (5.26) and (5.27). TP denotes the true positive rate (i.e. number of depressed (D) adolescents classified as depressed) and TN is the true negative rate (i.e. adolescents correctly identified as non-depressed (ND)). False positive (FP) is the number of ND adolescents incorrectly classified as D adolescents. False negative (FN) is the number of D incorrectly identified as ND.

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \quad (5.25)
\]

\[
\text{Sensitivity} = \frac{TP}{TP + FN} \times 100 \quad (5.26)
\]

\[
\text{Specificity} = \frac{TN}{TN + FP} \times 100 \quad (5.27)
\]

The aim of this research was to provide a method for a first stage mass-screening tool to assist psychologists with a full clinical evaluation of those screened as potentially depressed.

Therefore, it was more important to correctly identify depressed subjects (higher sensitivity) than to correctly detect controls (higher specificity). This is a common requirement made for diagnostic instruments in clinical research and is the assertion used in multiple depression detection studies [147][528].

This requires a high sensitivity to specificity ratio \((s_e/s_p > 1)\) without skewing \((s_e/s_p < 2)\). This is generally the intention although in some cases this is not achieved with high enough accuracy.
5.7 Experimental results on the Influence Model generation

5.7.1 Preliminary experiment using Auto and Cross Correlation (EXPprelim)

Psychological studies have used autocorrelation in as a representation of emotional inertia and showed significant differences between depressed and control adolescents \[87][88]\). An extension of concept is used here by using cross-correlation an investigation into not only the self-inertia (autocorrelation) but also cross-inertia between the parent and adolescent.

Cross-correlation is a measure of similarity of two series, \(f\) and \(g\), as a function of the lag, \(\tau\), of one relative to the other. In this application this would demonstrate a lagged relationship between the speakers (parent and adolescent) in a conversation with a measure of similarities in reference to the annotated sequence. Auto-correlation is related to the similarities at different lags within one sequence (speaker) in relation to annotated sequence.

The cross-correlation of continuous signals is given by equation (5.28) and for a discrete signal by equation (5.29) where \(f^*\) is the complex conjugate of sequence \(f\).

\[
(f*g)(\tau) = \int_{-\infty}^{\infty} f^*(t) g(t+\tau) dt \tag{5.28}
\]

\[
(f*g)[n] = \sum_{m=-\infty}^{\infty} f^*[m] g[m+n] \tag{5.29}
\]

In this application the constructs, generated from LIFE annotations, were merged into two states. In one setup only positive was retained and the remaining constructs (i.e. negative, neutral and silence) were contracted into a construct “not-positive” and the same for negative and “not-negative”. These were used to determine autocorrelation of positive and negative inertia up to a 10 second delay.

\[
S^i_t \in \{\text{Negative}(1), \text{NotNegative}(0)\}
\]

\[
S^i_t \in \{\text{Positive}(1), \text{NotPositive}(0)\}
\]
Figure 5-10 Relative difference between the non-depressed and depressed adolescent’s autocorrelation function for the negative (left) and positive (right) types. Including three interactions: event planning interaction (EPI), family consensus interaction (FCI) and problem solving interaction (PSI).

Figure 5-10 is the average relative difference between non-depressed and depressed autocorrelations for negative (left) and positive (right) annotation types. The autocorrelations demonstrate the similarities at different time delays with the adolescent annotated sequences. Negative and positive autocorrelations were higher for depressed compared to non-depressed in FCI and PSI. This supports evidence that those with poor psychological adjustment have higher levels of positive and negative emotional inertia (i.e. more resistance to emotional change) [88][87].

It has been shown that inertia is particularly noticeable in conflict tasks that have higher emotional demands including during disagreement (conflict tasks) or challenge (reminiscence tasks), as opposed to pleasurable events (positive tasks) [88]. The clear difference between the characteristics of the PSI and FCI compared to EPI could be based on the fact that the interactions were chosen to elicit different affect levels [96][97] with PSI setup to evoke situations to elicit conflicting behavior [276].
Figure 5-11 Relative difference between the non-depressed and depressed adolescent and parent sequence cross-correlation function for the negative (left) and positive (right) types. Including three interactions: event planning interaction (EPI), family consensus interaction (FCI) and problem solving interaction (PSI).

Figure 5-11 shows the average relative difference between non-depressed and depressed cross-correlation experiments using the negative (left) and positive (right) annotation types. The cross-correlations indicated the similarities at different time delays between the parent and adolescent annotated sequences. The results showed that the negative type cross-correlation was higher for depressed adolescents compared to the controls in all interactions. The positive type cross-correlation showed higher correlation in non-depressed adolescent conversations.

This indicates that depressed adolescents are more correlated to the past negative states of the parents and in contrast non-depressed adolescents are more correlated to the positive states of parents. This could agree with suggestions that the depressed adolescents are associated difficulties in regulation of negative emotions [137][24] and negative reciprocal parent-children process that reinforces the children’s depressive behaviors [137][135][144].

The results of auto- and cross-correlation of the dyadic annotated conversation sequences was a preliminary investigation for proof of concept in analyzing the annotations before moving on to a more sophisticated system.
5.7.2 Compare optimization methods of EIM (EXP1)

An experiment was carried out to find the ideal optimization method of the EIM using four construct states (positive, negative, neutral and silence). The six optimization methods (EM, SA, QN, MVR, GD and SA_GA) resulted in statistically identical parameters, based on ANOVA ($p>0.95$) comparing learned IC for each method.

To determine the optimization method to continue with further experiments the MSE and CPU time are supplied in Figure 5-12; averaged for a subset of the conversations. The MSE was almost constant, between 6.6-7.6%, within the six different optimization approaches. The EM was slightly less erroneous and additionally it was faster than the other optimization methods.

![Figure 5-12 Comparison of optimization methods, (Expectation Maximization (EM), Quasi-Newton (QN), Multivariate Linear Regression (MVR), Gradient Decent (GD), Simulated Annealing (SA) and Simulated Annealing/Genetic Algorithm Hybrid (SA/GA)), to learn the Emotional Influence Model (EIM) parameters using a delay of $t-1$. Performance is compared on the error rate (MSE) and CPU time (seconds) to learn the parameters, averaged over a portion each the ORI-DB.](image)

5.7.3 Determine effect of delay using Dynamic Influence Model (EXP2)

Based on the initial experimentation using cross-correlation it was observed that the reaction time was different for each speaker, which meant the lagged relationships were not consistent. Rather than using the commonly used time delay of one step, we have looked into the effect of multiple time delays as independent models.
The Log-likelihood (LL) was calculated over each subject separately, given by equation (5.8) as described in Section 5.3.2. Figure 5-13 shows the average LL over all ORI-DB LIFE annotations. It was revealed the DEIM had a maximum LL at \( t-1 \) (i.e. EIM), degraded up to \( t-5 \) and then tapered. This suggests shorter delays fit better, which is intuitive as prediction is easier from most recent states (\( t-1 \)). The longer delays still fit well and can provide insights in depression analysis and classification.

5.7.4 Determination of higher order dependence of EIM using HOEIM (EXP3)

The average Log-Likelihood, \( LL \), of constructed LIFE annotations comparing HOEIM implementations (MMM, ngram, KN-ngram) is given in Figure 5-13. The HOIEM outperform the DEIM with improved \( LL \) as the order increased for all HOEIM implementations; MMM was the worst, ngram better and KN smoothed n-gram the best. This suggested the past states are useful in conjunction with recent states and ties in with the results presented in Section 5.7.1 (\( Exp\)prelim) that indicated each participant had unique time lag associated with peak cross-correlation.

![Figure 5-13 Log-Likelihood (LL) comparison of different HOEIM implementations (MMM, ngram, KN-ngram) for each order (\( N=1,..,5 \)) and compared to the DEIM for delays of \( t-n=1,..,10 \)](image)

159
The HOEIM-MMM had the worst fit of all higher order approaches, yet it provided insights with separately weighted past delays where as the n-gram weights are equal. The average MMM weights, $\omega_{ij}(n_m)$, are shown over a subset of the ORI-DB illustrated by Figure 5-14.

It was observed that on average the MMM weight of self past states (self-transitions), $\omega_{ii}$, had the largest weighting for the most recent states and steadily dropped off with increased time lag. This would be interpreted as a speaker predominately being driven by their most recent emotions. This mirrors the results presented in Section 5.7.1 that indicated speakers had peak auto-correlation at short time delays.

In contrast the average MMM weight of past states of the other speaker (cross-transitions), $\omega_{ij}$, was approximately even across all delay lengths. The interpretation could be that there was an even spread of emotional influence from other speakers across multiple time delays. Another explanation could be that individual conversations could display peaks in weighting at many different delay lengths.

![Figure 5-14 Average MMM weights $\omega_{ij}(n_m)$ with time delay $n_m$ between the current and the previous state; $\omega_{11}(n_m)$–adolescent-adolescent, $\omega_{22}(n_m)$–parent-parent, $\omega_{12}(n_m)$–parent-adolescent, $\omega_{21}(n_m)$–adolescent-parent.](image)
Examination of the cross-transition MMM weights of individual conversations displayed a variation occurrence of the peak weight (i.e. \(\omega_{ij}(n_m)\) is maximum at a range of \(n_m\)). Such that each conversation has a differentiating lagged relationships.

The self-transition MMM weights showed the majority of conversations exhibited the same characteristics as the average. This confirmed that subjects are driven by their immediate past emotions (i.e. \(\omega_{ii}(n_m)\) is maximum at \(n_m = 1\)).

These concepts are illustrated by the examples of MMM weights given in Figure 5-15 for two different dyadic conversations. This suggests self-influence could be based on a shorter delay and cross-influence a larger weighting to further states.

![Figure 5-15 Mixed Memory Markov Model (MMM) weights \((\omega_{11}(n_m), \omega_{12}(n_m), \omega_{21}(n_m), \omega_{22}(n_m))\), with time delay \(n_m\) between the current and the previous state, generated for two different dyadic conversations (left and right).](image)

Table 5-3 quantifies the results providing statistics related to the time lag of the maximum MMM weights. The results showed for most cases the maximum self-transition weight occurred at \(t-1\) for 94% \((\omega_{11})\) and 97% \((\omega_{22})\) of the distribution. The peak time lag of cross-transition weights showed a distribution across lags of 8%-27% \((\omega_{12})\) and 6%-22% \((\omega_{21})\) on average occurring at \(t-3\) \((\omega_{12})\) and \(t-4\) \((\omega_{21})\).

<table>
<thead>
<tr>
<th>(\omega_{ij})</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>Average</th>
<th>Standard Deviation</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\omega_{11})</td>
<td>93.7</td>
<td>1.6</td>
<td>0.0</td>
<td>1.6</td>
<td>1.6</td>
<td>0.0</td>
<td>1.6</td>
<td>1.2</td>
<td>0.97</td>
<td>1</td>
</tr>
<tr>
<td>(\omega_{12})</td>
<td>17.5</td>
<td>27.0</td>
<td>7.9</td>
<td>9.2</td>
<td>11.1</td>
<td>15.9</td>
<td>11.1</td>
<td>3.6</td>
<td>2.10</td>
<td>3</td>
</tr>
<tr>
<td>(\omega_{21})</td>
<td>14.3</td>
<td>19.2</td>
<td>12.7</td>
<td>6.4</td>
<td>12.7</td>
<td>22.2</td>
<td>12.7</td>
<td>4.0</td>
<td>2.10</td>
<td>4</td>
</tr>
<tr>
<td>(\omega_{22})</td>
<td>96.8</td>
<td>0.0</td>
<td>1.6</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.1</td>
<td>0.28</td>
<td>1</td>
</tr>
</tbody>
</table>

161
5.8 Experimental Results of Influence model Parameters Analysis and Statistical

One of the objectives of this research was to create a conversation modeling system to determine risk factors that effect depressed adolescents. The features from EIM can compare the conversations for statistical differences and psychological interpretations.

5.8.1 Emotional Transition Probabilities Statistical Testing (EXP4)

To streamline only the first order (EIM) results are supplied for EPI, based on preliminary experiments using just 3 minutes of annotations. The self and cross transition probabilities were statistically examined using t-tests to compare the difference between depressed and control groups for each individual transition. \( p \)-values are given in Table 5-4, Table 5-5, Table 5-6 and Table 5-7 for \( P_{11}, P_{12}, P_{21}, \) and \( P_{22} \) emotional transition probabilities.

Shaded cells mark transitions that are on average more likely to be undertaken by depressed children or their parents compared to non-depressed. Clear cells indicate transitions more likely by non-depressed children or parents. Bold \( p \)-values indicate transitions with significant differences between depressed and non-depressed dyads.

### Table 5-4 T-test of Children’s Self-transition Probabilities (\( P_{11} \)) between Depressed and Control. Significant Transitions, \( p<0.05 \), are in Bold and Blue Shade Designate Transitions More Likely from Depressed

<table>
<thead>
<tr>
<th></th>
<th>(-)</th>
<th>(+)</th>
<th>(n)</th>
<th>(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(-)</td>
<td>0.240</td>
<td>0.267</td>
<td>0.940</td>
<td>0.812</td>
</tr>
<tr>
<td>(+)</td>
<td>0.506</td>
<td>0.030</td>
<td>0.889</td>
<td>0.112</td>
</tr>
<tr>
<td>(n)</td>
<td>0.006</td>
<td>0.862</td>
<td>0.264</td>
<td>0.064</td>
</tr>
<tr>
<td>(s)</td>
<td><strong>0.002</strong></td>
<td><strong>0.009</strong></td>
<td><strong>0.163</strong></td>
<td><strong>0.796</strong></td>
</tr>
</tbody>
</table>

Considering the statistically significant transitions in the EPI, Table 5-4, the following observations on the children’s behavior can be drawn:

- Indication non-depressed children are on average more likely to self-transition to positive from either or silence or positive, compared to non-depressed.
- The depressed children are significantly more likely to self-transition into a negative state from silence and from neutral.
The affect the parent has on the adolescent is represented by the $P_{12}$ transition probabilities, Table 5-5. Observations of the statistical tests are as follows:

- Depressed children are significantly more likely, compared to non-depressed counterparts, to transition into a negative state when their parent is negative.
- If the parent is silent, depressed adolescents are significantly more likely to be the group that turns to negative emotions.
- Non-depressed children are significantly more likely to transition into a positive state when their parent is in a negative state.

<table>
<thead>
<tr>
<th></th>
<th>(-)</th>
<th>(+)</th>
<th>(n)</th>
<th>(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(-)</td>
<td>0.001</td>
<td>0.011</td>
<td>0.082</td>
<td>0.444</td>
</tr>
<tr>
<td>(+)</td>
<td>0.177</td>
<td>0.101</td>
<td>0.319</td>
<td>0.369</td>
</tr>
<tr>
<td>(n)</td>
<td>0.476</td>
<td>0.514</td>
<td>0.247</td>
<td>0.315</td>
</tr>
<tr>
<td>(s)</td>
<td>0.037</td>
<td>0.106</td>
<td>0.203</td>
<td>0.518</td>
</tr>
</tbody>
</table>

The emotional affect the adolescent has on the parent is represented by the $P_{21}$ transition probabilities, Table 5-6; the only statistical significance is that:

- Parents of depressed children are significantly more likely than parents of non-depressed children to transition into silence when their child is silent.

<table>
<thead>
<tr>
<th></th>
<th>(-)</th>
<th>(+)</th>
<th>(n)</th>
<th>(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(-)</td>
<td>0.771</td>
<td>0.861</td>
<td>0.817</td>
<td>0.818</td>
</tr>
<tr>
<td>(+)</td>
<td>0.216</td>
<td>0.166</td>
<td>0.477</td>
<td>0.715</td>
</tr>
<tr>
<td>(n)</td>
<td>0.307</td>
<td>0.163</td>
<td>0.486</td>
<td>0.272</td>
</tr>
<tr>
<td>(s)</td>
<td>0.228</td>
<td>0.217</td>
<td>0.211</td>
<td>0.027</td>
</tr>
</tbody>
</table>
Table 5-7 T-TEST OF PARENT’S SELF-TRANSITION PROBABILITIES (P_{22}) BETWEEN DEPRESSED AND CONTROL. SIGNIFICANT TRANSITIONS, P<0.0, ARE IN BOLD AND BLUE SHAD DESIGNATE TRANSITIONS MORE LIKELY FROM PARENTS OF DEPRESSED CHILDREN

<table>
<thead>
<tr>
<th></th>
<th>(-)</th>
<th>(+)</th>
<th>(n)</th>
<th>(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(-)</td>
<td>0.073</td>
<td>0.072</td>
<td>0.691</td>
<td>0.694</td>
</tr>
<tr>
<td>(+)</td>
<td>0.342</td>
<td>0.409</td>
<td>0.971</td>
<td>0.314</td>
</tr>
<tr>
<td>(n)</td>
<td>0.636</td>
<td>0.533</td>
<td><strong>0.016</strong></td>
<td><strong>0.011</strong></td>
</tr>
<tr>
<td>(s)</td>
<td>0.222</td>
<td><strong>0.008</strong></td>
<td>0.830</td>
<td>0.126</td>
</tr>
</tbody>
</table>

The self-transitions of the parents are compared relating to the depressed state of their child. The observations based on the p-values in Table 5-7 are as follows:

- **a strong tendency** for parents of depressed children to transition from the neutral state into silent state.
- **Whereas, the parents of non-depressed children** are significantly more likely to self-transition from silence into a from neutral to neutral.
- **If the child is silent a parent transitions into a positive significantly more often if the adolescent is non-depressed.**

General observations comparing transition probabilities showed on average, depressed children are characterized by a strong tendency to transition into a negative state. Where as the positive state was more likely to be taken by a non-depressed adolescent compared to depressed. Depressed adolescents transition into negative behavior significantly more, even after no negative affect and could be that they perceived negative attention. The psychological interoperation, explained in the literature review, found depressed children incorrectly identify emotions and respond accordingly [319][320]. Parents of non-depressed children showed similar patterns transitioning to positive states. Parent of depressed children tend to negative states except with positive to positive maybe to support children.
5.8.2 Influence Coefficients for each EIM variation (EXP5)

5.8.2.1 DEIM Influence Coefficients and Discussion

The DEIM conducted for time delays \( n=1,2,\ldots,10 \) generated four contours representing a trajectory of IC as a function of time delay \( n, \theta_{ij}(n) \) \((i,j=1,2 \text{ and } n=1,2,\ldots,10)\), that estimate the extent a speaker was influenced from multiple delays.

IC reveals characteristics of extrinsic (cross-influences) and intrinsic (self-influences) processes responsible for monitoring, evaluating, and modifying emotional reactions. The intensity was expressed with IC and the trajectory describes temporal characteristics [501].

Figures 5-13, 5-14 and 5-16 show the average self-influence \((\theta_{ii})\) and cross-influence, \((\theta_{ij})\) of both the adolescent (left) and the parents (right). The average influence coefficients \((\theta)\) are supplied for the three interactions including event planning interaction (EPI), family consensus interaction (FCI) and problem solving interaction (PSI).

Each figure compares the set of average influence coefficient between adolescents with depression (D) and adolescent that are non-depressed (ND). This was also the setup for the parents with a comparison of average influence coefficients between the group of parents of depressed adolescents and parents of the non-depressed adolescents.
Figure 5-16 Average DEIM self-influence ($\theta_{ii}$ and $\theta_{22}$ (solid lines)) and cross-influence ($\theta_{12}$ and $\theta_{21}$ (dashed lines)) for time delays of $n=1,\ldots,10$, comparing parameters of depressed (D (blue line)) and non-depressed (ND (red line)) adolescents (left) and their parents (right) during the EPI.

The EPI IC trajectories for D and ND are illustrated by Figure 5-16. In descriptive terms, comparing groups the following observations can be made:

- The self-influence ($\theta_{ii}$) for the adolescents and the parents are on average higher for the non-depressed than the depressed group ($\theta_{ii}_{ND} > \theta_{ii}_{D}$)

- Conversely the cross-influence is higher for the adolescents that are depressed and higher for the parents of depressed adolescents ($\theta_{ij}_{ND} < \theta_{ij}_{D}$)

- From $t>8$ the cross influence of the parent to their child is greater than the self influence of the child ($\theta_{ij} > \theta_{ii}$) for the depressed group

- The parents self-influence are lower than adolescents average self-influence ($\theta_{22} < \theta_{11}$) so parents are more influenced by their child than child by the parent.
Figure 5-17 Average DEIM self-influence ($\theta_{11}$ and $\theta_{22}$ (solid lines)) and cross-influence ($\theta_{12}$ and $\theta_{21}$ (dashed lines)) for time delays of $n=1,\ldots,10$, comparing parameters of depressed (D (blue line)) and non-depressed (ND (red line)) adolescents (left) and their parents (right) during the FCI.

For the FCI, given by Figure 5-17 similar observations can be made to discriminate between the depressed and control groups. The changes to the relative IC values between the groups has begun to appear as the topics continues:

- For delays of $t\geq 6$ the depressed adolescents with depression have a lower self-influence compared to the non-depressed, ($\theta_{11, ND} > \theta_{11, D}$).

- In contrast, for $t<6$ the depressed adolescents have a higher self-influence compared to the average of the non-depressed group, ($\theta_{11, ND} < \theta_{11, D}$).

- Similar, parents of depressed adolescents present with higher self-influence for the shorter delays ($\theta_{22, ND} < \theta_{22, D}$) and for lower self-influence, ($\theta_{22, ND} > \theta_{22, D}$), for longer delay lengths, $n=8,9,10,\ldots$.

- Parents self-influence, for $n>3$, is lower than the adolescents self-influence ($\theta_{22} < \theta_{11}$) so parents are more influenced by their child than vice-versa.
Figure 5-18 Average DEIM self-influence ($\theta_{11}$ and $\theta_{22}$ (solid lines)) and cross-influence ($\theta_{12}$ and $\theta_{21}$ (dashed lines)) for time delays of $n=1,…,10$, comparing parameters of depressed (D (blue line)) and non-depressed (ND (red line)) adolescents (left) and their parents (right) during the PSI.

The IC trajectories for PSI are shown in Figure 5-18 and comparing differences between the emotional influence patterns of the depressed and non-depressed adolescents have the following observations:

- The self influence for every delay, except $t-1$, is higher for depressed than non-depressed adolescents ($\theta_{11,ND} < \theta_{11,D}$) which differs from EPI and FCI.

- The parents of non-depressed adolescents still have a higher self-influence compared to parents of depressed adolescents ($\theta_{22,ND} > \theta_{22,D}$) for most delays. Therefore the parents of depressed children are more influenced by their child.

- Again the parent has less self-influence than the child for the depressed group ($\theta_{22,D} < \theta_{11,D}$). In contrast, the non-depressed children in PSI are less self-influenced than the parents, ($\theta_{22,ND} > \theta_{11,ND}$).
Overall, there was consistency through all topics such that after approximately 5 seconds there was stabilization of the ICs up to a 10 second delay. This was consistent with results obtained by Kuppens et al. [88][87] that showed emotional inertia does not go beyond 5 seconds.

Additionally, the average values of self-ICs generally decreased with an increase in the time delay. This is an intuitively expected pattern, where both parents and adolescents are predominantly driven by their own emotions ($\theta_{11}(n) > \theta_{12}(n)$) and $\theta_{22}(n) > \theta_{21}(n)$ for all $n$). This supported the auto-correlation results having a high correlation with recent self-past states.

It was evident that the self-ICs are mostly higher for the adolescents compared to their parents. Subsequently parents have higher cross-ICs and are emotionally more influenced by children than children by parents. This could be viewed as parents being more attuned to the behavior of their children and could explain why the difference is more prominent in the depressed.

This is explained in psychological literature, as the parent’s evolutionary adaption to understand emotional needs [116]. This was also found in the case of child interacting with psychologists attuning to the child’s behavioral cues, deliberately or spontaneously, based on impairments [520][519][521].
In descriptive terms, a summary of observations in relation to depression can be derived from the Influence Coefficient’s trajectories illustrated the figures:

1. In EPI adolescents and their parents showed a higher self-influence on average for the non-depressed group compared to depressed ($\theta_{ii,ND}(n) > \theta_{ii,D}(n)$)

2. Where as FCI topic the adolescents had a higher self influence for the non-depressed only for $n>5$ and the parent $n>7$. For shorter delay lengths the depressed had higher self-influence in adolescents

3. Depressed adolescents on average displayed higher self-influence compared to non-depressed in the PSI. The parent still displayed higher self influence for the non-depressed so they are more influence by their child ($\theta_{ij,ND}(n) < \theta_{ij,D}(n)$)

The topic plays an important role; evident by the fact the class (depressed or non-depressed) with higher self-influences varies between topics. A direct comparison of topics for adolescent’s self-ICs is given in Figure 5-19 for the control and depressed groups. The graphs show the ICs were considerably higher for the PSI session in the depressed group compared to EPI and to a lesser extent FCI. In contrast the non-depressed adolescents had a higher self-influence in EPI and low in PSI.
Figure 5-19 DEIM generated self Influence Coefficients (IC), $\theta_{ii}$, for each time delay, $t-n$, comparing the self IC for each interaction (EPI, FCI, PSI) given for the non-depressed adolescents, $\theta_{11_{ND}}$, (left) and depressed adolescents, $\theta_{11_{D}}$, (right)

The event planning (EPI) was the first ‘warm-up’ conversation not indented to have many altercations. As the topics follow in to the family consensus and problem solving more issues would arise between the interaction families such as conflict.

The PSI sessions are more likely to evoke conflictual behavior, which have strong correlation with adolescent depression in family environments [276]. Emotional demands may be higher in tasks that have disagreement (conflict tasks) or challenge (reminiscence tasks), as opposed to pleasurable events (positive tasks).

This could provide an explanation as to why there was so much self-influence in PSI for depressed adolescents, compared to other topics, and largest difference to parent self-influence. Moreover, this could explain why depressed, compared to non-depressed, have a higher self-influence in PSI but there reverse for EPI.

Kuppens et al. showed that those suffering from psychological maladjustment are more resistant to change and thus display higher inertia, shown through the high self-ICs [88][87]. It was found that differences in inertia may be particularly salient in conflict tasks and emotionally evocative interactions [88] and results from the DEIM (EXP5) and the correlation (EXPprelim) experiments support this evidence.
5.8.2.2 DEIM Influence Coefficients Statistical Tests

For a statistical comparison of the two independent variables multivariate analysis of variance (MANOVA) was conducted on pairwise comparison of depressed and controls. MANOVA tests multiple dependent variables, in this case IC, at once to determine if there is correlation to the independent variable (depressed or control).

The Wilk’s lambda test estimated differences between the means of two classes where significant level of correlation to depression was considered with a \( p \) value less than 0.05.

Table 5-8 Statistical Analysis of the Entire Set of DEIM Generated Influence Coefficients (\( \theta_{11}, \theta_{12}, \theta_{21}, \theta_{22} \)) Comparing Between Depressed and Non-Depressed Adolescents. Results Give the MANOVA Wilks Lambda P-value, Where Significant (\( p<0.05 \)) Parameters Are Shaded, for Each Topic (EPI, FCI, PSI and ALL Topics) and DEIM Delay (\( t-1,\ldots,t-10 \)).

<table>
<thead>
<tr>
<th>t-n</th>
<th>EPI</th>
<th>FCI</th>
<th>PSI</th>
<th>ALL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.001</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0.137</td>
<td>0.003</td>
<td>0.01</td>
<td>0.05</td>
</tr>
<tr>
<td>3</td>
<td>0.012</td>
<td>0.159</td>
<td>0.04</td>
<td>0.044</td>
</tr>
<tr>
<td>4</td>
<td>0.021</td>
<td>0.062</td>
<td>0.005</td>
<td>0.02</td>
</tr>
<tr>
<td>5</td>
<td>0.09</td>
<td>0.434</td>
<td>0.007</td>
<td>0.278</td>
</tr>
<tr>
<td>6</td>
<td>0.046</td>
<td>0.063</td>
<td>0.147</td>
<td>0.137</td>
</tr>
<tr>
<td>7</td>
<td>0.041</td>
<td>0.124</td>
<td>0.26</td>
<td>0.442</td>
</tr>
<tr>
<td>8</td>
<td>0.015</td>
<td>0.26</td>
<td>0.114</td>
<td>0.027</td>
</tr>
<tr>
<td>9</td>
<td>0.004</td>
<td>0.164</td>
<td>0.017</td>
<td>0.273</td>
</tr>
<tr>
<td>10</td>
<td>0.213</td>
<td>0.069</td>
<td>0.005</td>
<td>0.608</td>
</tr>
<tr>
<td>All</td>
<td>0.128</td>
<td>0.007</td>
<td>0.001</td>
<td>0</td>
</tr>
</tbody>
</table>

The MANOVA results, in Table 5-8, shows for most delays a combined set of IC had significant difference between the two groups. Individually each DEIM delay was useful in depression analysis and combining was not necessarily better.
The parameters that are significant can provide psychological interpretations to determine qualitative factors that affect depression. Determining that the extracted features have significant differences they are useful for training a classification system to automatically diagnose depression from an unknown test sample.

Following MANOVA analysis, individual coefficients were compared using univariate analysis. Pair-wise t-tests, given in Table 5-9, on the ICs have multiple significant (\(p<0.05\)) coefficients. The self and cross ICs (\(\theta_{ii}\) and \(\theta_{ij}\)) are constrained to sum to one and are therefore inversely proportional and \(p\) values are identical.

### 5.8.2.3 HOEIM Influence Coefficients and Discussion

The HOEIM was created for \(N=1,2,\ldots,5\), where each order considers delay steps of \(n=1,\ldots,N\). The HOEIM ICs are illustrated in Figure 5-20, Figure 5-21 and Figure 5-22 for the EPI, FCI and PSI separately for each dyad. The general trends are the same as the DEIM; the self-influence decreases as the order increases. The reduction was less evident in the HOEIM due to the first delay contributing in the higher orders.

---

<table>
<thead>
<tr>
<th>t-n</th>
<th>EPI</th>
<th>FCI</th>
<th>PSI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(\theta_{11})</td>
<td>(\theta_{22})</td>
<td>(\theta_{12})</td>
</tr>
<tr>
<td>1</td>
<td>0.262</td>
<td>0.002</td>
<td>0.002</td>
</tr>
<tr>
<td>2</td>
<td>0.098</td>
<td>0.098</td>
<td>0.098</td>
</tr>
<tr>
<td>3</td>
<td>0.027</td>
<td>0.215</td>
<td>0.027</td>
</tr>
<tr>
<td>4</td>
<td>0.122</td>
<td>0.122</td>
<td>0.122</td>
</tr>
<tr>
<td>5</td>
<td>0.038</td>
<td>0.547</td>
<td>0.038</td>
</tr>
<tr>
<td>6</td>
<td>0.03</td>
<td>0.23</td>
<td>0.23</td>
</tr>
<tr>
<td>7</td>
<td>0.126</td>
<td>0.091</td>
<td>0.126</td>
</tr>
<tr>
<td>8</td>
<td>0.026</td>
<td>0.097</td>
<td>0.097</td>
</tr>
<tr>
<td>9</td>
<td>0.013</td>
<td>0.063</td>
<td>0.013</td>
</tr>
<tr>
<td>10</td>
<td>0.09</td>
<td>0.417</td>
<td>0.09</td>
</tr>
</tbody>
</table>
In descriptive terms a comparison of the depressed and non-depressed classes the following observations are made in reference to the EPI:

- The self-influence for the adolescents and the parents are on average lower for the depressed than the non-depressed group for the EPI ($\theta_{ii,ND} > \theta_{ii,D}$)

- Conversely the cross-influence is higher for the adolescents that are depressed and higher for the parents of depressed adolescents ($\theta_{ij,D} > \theta_{ij,ND}$)

- The parents self-influence are lower than adolescents average self-influence, ($\theta_i > \theta_{ij}$) so parents are more influenced by their child than vice-versa.

These results indicated that during EPI, both depressed adolescents and their parents were more strongly driven by reactions to their own emotional states than the non-depressed group. For both groups, the adolescents and parents were predominantly driven by reactions to their own previous emotional states.
Figure 5-21 Average HOEIM self-influence ($\theta_{11}$ and $\theta_{22}$ (solid lines)) and cross-influence ($\theta_{12}$ and $\theta_{21}$ (dashed lines)) for orders $N=1,\ldots,5$, comparing parameters of depressed (D (blue line)) and non-depressed (ND (red line)) adolescents (left) and their parents (right) during the FCI.

For the FCI topic observations can be made to discriminate between the depressed and control groups as follows:

- In contrast to the EPI, the self-influences of the depressed adolescents and their parents (except 1st order) were higher compared to the non-depressed adolescents or parent ($\theta_{ii,D} > \theta_{ii,ND}$)

- Conversely the cross-influences are of the parents and adolescents are lower in the depressed group compared to the non-depressed adolescents ($\theta_{ij,D} < \theta_{ij,ND}$)

- The parent’s self-influence is mostly lower than adolescent’s self-influence, ($\theta_{ii} > \theta_{ij}$). Such that parents are more influenced by their child as with the EPI.

These results indicated that during FCI, differences between the depressed and non-depressed groups were generally reversed compared with EPI. This means that compared with the non-depressed group, the depressed adolescents and their parents were more strongly driven by reactions to their own previous emotional states, and less by reactions to the emotional states of their conversational partners.
Comparing emotional influence patterns between the depressed and non-depressed adolescents during the PSI have the following observations:

- As in the EPI, the self influence is lower for parents of depressed adolescents compared to parents of the non-depressed adolescents ($\theta_{ii,D} > \theta_{ii,ND}$)

- Similar to FCI, the adolescents with depression have higher self-influence compared to the non-depressed group ($\theta_{ii,ND} > \theta_{ii,D}$)

- Again the parent has a lower self-influence compared to the child; but only for the depressed group ($\theta_{ii,D} > \theta_{ij,D}$). In contrast, in the PSI the non-depressed children are less self-influenced compared to their parents ($\theta_{ij,ND} > \theta_{ii,ND}$)

These results indicated that during PSI the depressed adolescents, compared with non-depressed, were more strongly driven by self-influences. As indicated in [504], this particular type of behavior could be a sign of a stronger desire or struggle by the depressed adolescents to achieve emotional autonomy from their parents [505]. This could explain why it is more prominent in PSI due to being set up to be more conflictual than during EPI and FCI.
An overall summary of observations made from the HOEIM parameters in terms comparing influences between depressed and control groups is as follows:

1. Both parents and adolescents are predominantly driven by their own influences (intra-speaker influences are higher than inter-speaker influences)

2. As the order increases, the cross influence increases and self-influence reduces and could be interpreted as speakers becoming attuned to emotions conveyed by their conversational partners as time delay increases. A person dependent time lag is needed to observe speakers’ reactions. These observations show consistency with previously reported existence of “emotional inertia” [88][87].

3. The parents’ self-influences are lower than the self-influences of their adolescent children. The parents are therefore more influenced by the adolescent than the adolescent is by the parents. This emotional behavior shows parents appear to be attuned to others affective behavior, reflecting evolution in high parental sensitivity to offspring needs [116].

4. Additionally it is observed that on average, compared to the controls, the depressed adolescents have a higher self-influence for the argumentative topics (FCI and PSI).

Previous studies have shown depressed adolescents have higher emotional inertia, analogous to the IC [88][87]. HOEIM supports this evidence with depressed adolescents having higher self-influence for conflictive topics (FCI and PSI). It is evident that depressed adolescents largely, have a higher self-influence compared to parents. This is equivalent to parents being more influenced by their child than vice-versa and most evident in depressed, especially during PSI. The psychological interoperation could be parents attune to the child as an evolutionary adaption [116].
Again, as with DEIM, it can be concluded topic plays a role in conversation emotional patterns. The dominating group in both the parents and children trends was consistent between the DEIM and HOEIM. Explicitly, a comparison of the children’s self-influence for each interaction is given in Figure 5-23 for the depressed and control groups and shows EPI had most self-influence in ND and PSI in the D group.

5.8.2.4 HOEIM Influence Coefficients Statistical tests

<table>
<thead>
<tr>
<th>Order (N)</th>
<th>EPI</th>
<th>FCI</th>
<th>PSI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0001</td>
<td>0.0014</td>
<td>0.0001</td>
</tr>
<tr>
<td>2</td>
<td>0.0006</td>
<td>0.0006</td>
<td>0.0042</td>
</tr>
<tr>
<td>3</td>
<td>0.0048</td>
<td>0.0256</td>
<td>0.0044</td>
</tr>
<tr>
<td>4</td>
<td>0.0068</td>
<td>0.0005</td>
<td>0.0023</td>
</tr>
<tr>
<td>5</td>
<td>0.0053</td>
<td>0.0073</td>
<td>0.0280</td>
</tr>
</tbody>
</table>

MANOVA was used to determine the ICs of the HOEIM that are statistically significant for each of the topics. Results of the MANOVA test on each HOEIM order is given in Table 5-10 using the ICs for each topic independently. All tests confirmed statistically significant ICs differentiating between the depressed and control groups.
5.9 Experimental Results of Depression Classification

The features learned from the EIM can be considered as a conversational signature \( \{\Phi, P\} \), where \( \Phi \) is the set of IC \((\theta_{11}, \theta_{12}, \theta_{21}, \theta_{22})\) and \( P \) is the set of transition probabilities \((P_{11}, P_{12}, P_{21}, P_{22})\). The features can be used in a classification system to objectively detect adolescent depression based on emotional influence patterns.

In the DEIM case three classification methods have been compared as outlined in Section 5.4.2: Support Vector Machine (SVM), Neural Network (NN) and improvement on the NN using optimization on the number of hidden nodes (O-NN).

In the HOEIM case a comparison of implementations (MMM and KN-ngram) with SVM classification led to 2-stage mRMR feature optimization.

5.9.1 Depression detection using the DEIM with SVM (EXP6)

SVM depression classification performance using the DEIM features, \( \{\Phi, P\} \), is evaluated by accuracy, sensitivity and specificity in Table 5-11. The average accuracy for all delays is 60.8%, 65.6% and 80.0% for EPI, FCI and PSI. For EPI the best delay was \( t-9 \) (64.47%) and worst \( t-4 \) (58.50%). FCI ranged between \( t-1 \) (68.16%) and \( t-10 \) (63.15%). PSI was the best overall between 74.13% (\( t-3 \)) and 66.49% (\( t-10 \)).

<table>
<thead>
<tr>
<th>DEIM Delay</th>
<th>EPI</th>
<th>FCI</th>
<th>PSI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc</td>
<td>Sens</td>
<td>Spec</td>
</tr>
<tr>
<td>1</td>
<td>60.3</td>
<td>55.2</td>
<td>64.7</td>
</tr>
<tr>
<td>2</td>
<td>62.1</td>
<td>51.9</td>
<td>70.8</td>
</tr>
<tr>
<td>3</td>
<td>59.3</td>
<td>48.9</td>
<td>48.9</td>
</tr>
<tr>
<td>4</td>
<td>58.5</td>
<td>47.9</td>
<td>67.5</td>
</tr>
<tr>
<td>5</td>
<td>61.4</td>
<td>50.2</td>
<td>71.0</td>
</tr>
<tr>
<td>6</td>
<td>61.4</td>
<td>46.9</td>
<td>73.7</td>
</tr>
<tr>
<td>7</td>
<td>60.3</td>
<td>49.0</td>
<td>69.7</td>
</tr>
<tr>
<td>8</td>
<td>61.7</td>
<td>49.4</td>
<td>72.3</td>
</tr>
<tr>
<td>9</td>
<td>64.5</td>
<td>53.2</td>
<td>74.1</td>
</tr>
<tr>
<td>10</td>
<td>58.6</td>
<td>46.0</td>
<td>69.3</td>
</tr>
</tbody>
</table>

Table 5-11 Depression Classification Performance using a SVM and the Entire DEIM Feature Set \((P_{11}, P_{12}, P_{21}, P_{22}, \theta_{11}, \theta_{12}, \theta_{21}, \theta_{22})\) for Each Delay and Interaction
For most DEIM setups EPI has the lowest accuracy and PSI detects depression at the highest average rate with optimal sensitivity, confirming past studies findings using the same interactions [140][454]. This could be attributed to the fact conversations were chosen to elicit different affect levels in each topic [96][97] and PSI setup to evoke situations to elicit conflicting behavior [276]. This could result in more significant emotional transitions and influences alterations in depression.

5.9.2 Effectiveness of DEIM for depression classification using NN (EXP7)

The previous experiment, EXP6, was repeated with NN classification of DEIM features given by the results Table 5-12 for all delay lengths. No trend between delay and depression classification accuracy was found other than t-1 is never the optimal delay length and performance was generally optimum around the moderate delay length.

Comparing topics revealed PSI was better than FCI for the majority of delays by an average of 6%. PSI was only better than EPI for shorter delay lengths by an average of 18% but for the longer delays EPI was on average 14% more accurate than PSI.

<table>
<thead>
<tr>
<th>DEIM delay</th>
<th>Feature Set: P_{11}, P_{12}, P_{21}, P_{22}, θ_{11}, θ_{12}, θ_{21}, θ_{22}</th>
<th>DEIM parameters</th>
<th>Classifier: NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>EPI</td>
<td>FCI</td>
<td>PSI</td>
<td></td>
</tr>
<tr>
<td>Acc</td>
<td>Sens</td>
<td>Spec</td>
<td>Acc</td>
</tr>
<tr>
<td>1</td>
<td>78.6</td>
<td>76.7</td>
<td>77.8</td>
</tr>
<tr>
<td>2</td>
<td>53.2</td>
<td>51.0</td>
<td>52.0</td>
</tr>
<tr>
<td>3</td>
<td>82.1</td>
<td>77.7</td>
<td>96.7</td>
</tr>
<tr>
<td>4</td>
<td>92.8</td>
<td>88.3</td>
<td>96.0</td>
</tr>
<tr>
<td>5</td>
<td>95.0</td>
<td>96.1</td>
<td>94.3</td>
</tr>
<tr>
<td>6</td>
<td>95.0</td>
<td>96.1</td>
<td>94.3</td>
</tr>
<tr>
<td>7</td>
<td>77.8</td>
<td>75.0</td>
<td>79.4</td>
</tr>
<tr>
<td>8</td>
<td>95.0</td>
<td>96.1</td>
<td>94.3</td>
</tr>
<tr>
<td>9</td>
<td>92.5</td>
<td>90.0</td>
<td>95.0</td>
</tr>
<tr>
<td>10</td>
<td>76.7</td>
<td>73.0</td>
<td>79.8</td>
</tr>
</tbody>
</table>
Compared to the SVM the NN classification system improved depression detection in adolescents by an average of 23.1%, 9.7% and 10.4% for EPI, FCI and PSI. The topic and DEIM delay performance differs between SVM and NN, which indicated a bias and importance of classification system.

5.9.3 Optimized NN for Depression Classification using the DEIM (EXP8)

The number of hidden nodes greatly affects NN performance so to ensure the best results a CV procedure was used to find the optimum nodes, described in Section 5.4.2.3, testing NN with 10, 15, 20, 25, 30, 35, 40 and 50 hidden nodes. The average results, across the delays, are presented as a function of hidden nodes versus accuracy of classification in Figure 5-24. The EPI and PSI topic peaks at 25 nodes but the FCI topic, which struggled in the 25 nodes NN, has a less obvious hidden node size.

Figure 5-24 Average accuracy of NN classification with respect to the number of hidden nodes. Results averaged over delay lengths \( n=1, 2, \ldots, 10 \) separately for each of the interactions (EPI, FCI, PSI) and each node iteration (10, 15, 20, 25, 30, 35 and 40)
The optimum number of hidden nodes is given in Table 5-13 for each delay length and corresponding accuracies in Table 5-14 using optimum nodes. The results indicated the 25 nodes are optimal for most setups but other sizes did improve results.

Table 5-13 Optimum Number of Hidden Nodes in a NN for the Best Performance in Depression Classification Using 5-fold CV Given for Each DEIM Feature Set and for Each Interaction (EPI, FCI, PSI)

<table>
<thead>
<tr>
<th>Optimized Number of Hidden Nodes</th>
<th>DEIM(n)</th>
<th>EPI</th>
<th>FCI</th>
<th>PSI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>30</td>
<td>10</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>30</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>15</td>
<td>30</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>25</td>
<td>40</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>25</td>
<td>20</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>20</td>
<td>30</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>25</td>
<td>25</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>25</td>
<td>20</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>25</td>
<td>20</td>
<td></td>
</tr>
</tbody>
</table>

The results suggest accuracy varies across delays and it was clear that delays other than $t-1$ were better. The best result with the O-NN for FCI was with a delay of $n=5$ reaching 97.5%, for EPI a slightly lower top accuracy at 95% for multiple delays $n=5, 6$ and 8 and PSI the lowest top result, $t-2$, but still reached 93%. The average accuracy across all delays was 88%, 88% and 84% for EPI, FCI and PSI.

Table 5-14 Depression Classification Performance Using the Optimized Hidden Node Neural Network (O-NN) with the Entire DEIM Feature Set ($P_{11}, P_{12}, P_{21}, P_{22}, \theta_{11}, \theta_{12}, \theta_{21}, \theta_{22}$) for Each Delay and Interaction

<table>
<thead>
<tr>
<th>Feature Set: $P_{11}, P_{12}, P_{21}, P_{22}, \theta_{11}, \theta_{12}, \theta_{21}, \theta_{22}$ DEIM parameters</th>
<th>Classifier: Optimized NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEIM Delay</td>
<td>EPI</td>
</tr>
<tr>
<td>Acc</td>
<td>Sens</td>
</tr>
<tr>
<td>1</td>
<td>79.2</td>
</tr>
<tr>
<td>2</td>
<td>84.3</td>
</tr>
<tr>
<td>3</td>
<td>81.9</td>
</tr>
<tr>
<td>4</td>
<td>92.8</td>
</tr>
<tr>
<td>5</td>
<td>95.0</td>
</tr>
<tr>
<td>6</td>
<td>95.0</td>
</tr>
<tr>
<td>7</td>
<td>78.9</td>
</tr>
<tr>
<td>8</td>
<td>95.0</td>
</tr>
<tr>
<td>9</td>
<td>92.5</td>
</tr>
<tr>
<td>10</td>
<td>89.3</td>
</tr>
</tbody>
</table>
5.9.4 Compare the DEIM classification system for depression detection

The average accuracy, across DEIM delays, is given for all topics and classifiers (SVM, NN, O-NN) in Table 5-15. The O-NN improved the performance compared to all SVM and the 25 node NN. The O-NN improved on the NN by an average of approximately 7% for the three interactions and delays. The SVM was clearly the worst performing classifier on average 21% below the O-NN and 14% below the NN.

Table 5-15 COMPARISON OF CLASSIFICATION METHODS (O-NN, NN, SVM) USING THE ENTIRE SET OF DEIM PARAMETERS ($P_{11}$, $P_{12}$, $P_{21}$, $P_{22}$, $\theta_{11}$, $\theta_{12}$, $\theta_{21}$, $\theta_{22}$) AS A FEATURE SET. RESULTS ARE GIVEN AS AN AVERAGE ACCURACY (%) OF ALL TEN DEIM DELAYS (T-1, T-2, ..., T-10)

| Feature Set: $P_{11}$, $P_{12}$, $P_{21}$, $P_{22}$, $\theta_{11}$, $\theta_{12}$, $\theta_{21}$, $\theta_{22}$ DEIM parameters | Average Accuracy (%) for all DEIM Delays |
|---|---|---|
| Classifier | EPI | FCI | PSI |
| ONN | 88.4 | 88.3 | 84.2 |
| NN | 83.9 | 75.3 | 81.3 |
| SVM | 60.8 | 65.6 | 71.0 |

Figure 5-25 shows the average accuracy, for three topics, for each classifier and DEIM delay. The SVM was generally the worst and the O-NN the best classification method. The highest average accuracy achieved was using DEIM features at $t-7$, with 93%, using the O-NN, NN still reached a peak of 91% ($t-4$) and SVM 67.1% ($t-5$).

Figure 5-25 Comparison of average depression classification accuracy using the DEIM features ($P_{11}$, $P_{12}$, $P_{21}$, $P_{22}$, $\theta_{11}$, $\theta_{12}$, $\theta_{21}$, $\theta_{22}$) at each delay ($t-n$) for the SVM, NN and Optimized NN
The receiver operating characteristics (ROC), Figure 5-26, are used to evaluate the performance of three depression classification methods (SVM, NN, O-NN) using DEIM feature set. The ROC space determines the combination of topic, order and classifier that achieves the best true positive rate to false positive rate ratio. The data point closest to the top left corner of the ROC space signifies this best setup.

Figure 5-26 ROC space of depression classification comparing three classification models SVM (red), 25 node Neural Network (blue) and optimized node NN (green) for EPI (left), FCI (middle) and PSI (right). Each point on the ROC space corresponds to the set time delay \( n=1,\ldots,10 \) used to extract the DEIM feature set.

Within each interaction the SVM performance of the different delays of the DEIM show little variation. This was not true with NN and O-NN, which showed larger depression classification accuracy variation between delays lengths. Generally the sensitivity and specificity ratio was best in the PSI and O-NN compared to SVM. It should be noted the O-NN experimental results might be optimistic as the parameters were tuned during preliminary experiments.
5.9.5 Effectiveness of HOEIM parameters for depression classification (EXP9)

Similar as the DEIM classification system the HOEIM features can be used detect depression. The main complication in this machine-learning problem was the exponentially increasing feature set. The larger dimensional feature set can be an issue and require more hidden nodes and lead to inefficient expensive O-NN.

The SVM is theoretically better as it is robust with high-dimensional data and the comparatively limited samples in this application. SVM can handle high-dimensional data, does not suffer from the curse of dimensionality, avoids over fitting and optimization is computational efficient [61][407][411].

5.9.5.1 Depression detection using HOEIM-MMM parameters

SVM depression classification performance with HOEIM parameters implemented with MMM is shown in Table 5-16. The accuracy of HOEIM-MMM parameters improved to the 3rd order and then decreased. Compared to the 1st order the 2nd, 3rd, 4th and 5th improved by an average accuracy of 5.7%, 7.9%, 3.5% and 0.5%. The highest accuracy attained was 65.1% (2nd), 80.2% (3rd) and 81.2% (3rd) for EPI, FCI and PSI.

It is noted for HOEIM-MMM depression classification, PSI was the best performing interaction, FCI slightly lower and EPI considerably worse. Across the orders the average accuracy for the EPI, FCI and PSI is 60.9%, 75.3% and 75.5%.

Table 5-16 Depression Classification Performance using SVM with the Entire Set of HOEIM-MMM Features \((P_{11}, P_{12}, P_{21}, P_{22}, \theta_{11}, \theta_{12}, \theta_{21}, \theta_{22})\) for Each Order (N) and Interaction

| Feature Set: \(P_{11}, P_{12}, P_{21}, P_{22}, \theta_{11}, \theta_{12}, \theta_{21}, \theta_{22}\) HOEIM-MMM parameters | Classifier: SVM |
|---|---|---|---|---|---|---|---|
| Order (N) | EPI | FCI | PSI |
| | Acc | Sens | Spec | Acc | Sens | Spec | Acc | Sens | Spec |
| 1 | 60.3 | 55.2 | 64.7 | **68.2** | **67.7** | **68.5** | 72.8 | 75.9 | 70.1 |
| 2 | 65.1 | 58.6 | 70.6 | 79.8 | 90.8 | 67.8 | 73.3 | 80.0 | 66.4 |
| 3 | 63.5 | 62.1 | 64.7 | 80.2 | 91.0 | 69.1 | 81.2 | 88.3 | 73.6 |
| 4 | 58.7 | 55.2 | 61.8 | 76.1 | 86.7 | 66.2 | 76.9 | 78.0 | 75.6 |
| 5 | 57.1 | 58.6 | 55.9 | 72.1 | 79.0 | 65.3 | 73.5 | 76.6 | 70.7 |
5.9.5.2 Depression detection using HOEIM-KN-ngram parameters

The depression classification performance of the SVM using the HOEIM parameters implemented with KN-ngram to estimate the higher order model is shown in Table 5-17. Generally, the 1\textsuperscript{st} order was the among lowest performing with the higher order HOEIM on average, across topics, improved by 5\%, 6.8\%, 5.9\% and 6.5\% for the 2\textsuperscript{nd}, 3\textsuperscript{rd}, 4\textsuperscript{th} and 5\textsuperscript{th} orders respectively. The highest accuracy and HOEIM-KN-ngram order was 69.8\% (2\textsuperscript{nd} and 5\textsuperscript{th}), 83.4\% (3\textsuperscript{rd}) and 82.6\% (3\textsuperscript{rd}) for EPI, FCI and PSI.

Table 5-17 Depression Classification Performance using SVM with the Entire Set of KN-ngram HOEIM Features \((P_{11}, P_{12}, P_{21}, P_{22}, \theta_{11}, \theta_{12}, \theta_{21}, \theta_{22})\) for Each Order (N) and Interaction

<table>
<thead>
<tr>
<th>Order (N)</th>
<th>EPI Acc</th>
<th>Sens</th>
<th>Spec</th>
<th>FCI Acc</th>
<th>Sens</th>
<th>Spec</th>
<th>PSI Acc</th>
<th>Sens</th>
<th>Spec</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>61.9</td>
<td>55.2</td>
<td>67.7</td>
<td>70.9</td>
<td>74.0</td>
<td>67.4</td>
<td>77.2</td>
<td>82.7</td>
<td>70.8</td>
</tr>
<tr>
<td>2</td>
<td>69.8</td>
<td>62.1</td>
<td>76.5</td>
<td>81.5</td>
<td>89.7</td>
<td>71.9</td>
<td>73.5</td>
<td>81.3</td>
<td>64.3</td>
</tr>
<tr>
<td>3</td>
<td>64.3</td>
<td>69.0</td>
<td>59.3</td>
<td>83.4</td>
<td>91.6</td>
<td>73.7</td>
<td>82.6</td>
<td>89.3</td>
<td>74.7</td>
</tr>
<tr>
<td>4</td>
<td>68.3</td>
<td>65.5</td>
<td>70.6</td>
<td>79.5</td>
<td>85.9</td>
<td>72.0</td>
<td>79.8</td>
<td>86.4</td>
<td>72.0</td>
</tr>
<tr>
<td>5</td>
<td>69.8</td>
<td>65.5</td>
<td>73.5</td>
<td>80.5</td>
<td>87.0</td>
<td>73.0</td>
<td>79.3</td>
<td>85.5</td>
<td>71.9</td>
</tr>
</tbody>
</table>

It was observed that the HOEIM-KN-ngram parameters used in SVM adolescent depression detection was worst in the EPI. This is consistent with observations from DEIM SVM depression classification.

The PSI and FCI detected depression in adolescents effectively the same average rate, each the best for certain orders. The average accuracy for all orders was 66.8\%, 79.2\% and 78.5\% for the EPI, FCI and PSI.
5.9.6 Depression detection using optimized EIM parameters (EXP10)

The conversational signature \( \{\Phi, P\} \) was a very large feature set, especially as the HOEIM order increases and more so for KN-ngram than MMM. In general not all of the individual components \( \{\Phi, P\} \) best discriminate between depressed and control subjects with many features redundant or irrelevant. It is valuable to reduce data-dimensionality and produce an optimized sub-set using a feature selection strategy.

In this application Minimum redundancy and maximum relevancy (mRMR) was implemented as a fast, powerful and efficient filter approach to generate a sub-set of optimal features. The feature sub-set generated by mRMR best characterizes properties of the target subject to the constraint that the features are mutually dissimilar[37].

![Diagram](image.png)

Figure 5-27 Framework of the feature selection method using a 2-stage mRMR filter to generate a ranking of top acoustic feature category used in a SVM wrapper stage to iteratively remove the lowest ranked feature to find the optimal accuracy

The relevance and redundancy calculation can be computed using correlation, t-test, F-test or distances. Commonly the mRMR is defined through Mutual Information Quotient. The MIQ is beneficial as it finds features that jointly have the maximum statistical "dependency" on the classification variable [38].

The MIQ is defined by \((5.30)\) where \(C\) is the class (D or ND) and \(i\) define the current index of the selected feature, \(j\) is a feature that already belongs to the optimal subset given by \(S\). From this \(I(i,C)\) is the mutual information between feature \(i\) and class \(C\) and \(I(i,j)\) is the mutual information between features \(i\) and \(j\).
The mutual information between two features is given by equation (5.31), where \( p(x) \) or \( p(y) \) is the probability density function of variable \( x \) or \( y \), the joint probability density function between \( x \) and \( y \) is \( p(x,y) \) and \( i \) and \( j \) denote the indices.

It has been suggested to follow up after the filter mRMR method with a second-stage wrapper method to improve results [38]. The wrapper stage optimized the feature set based on maximizing the 5-fold CV classification accuracy of SVM by iteratively removing the lowest rank features from the mRMR optimized subset. It should be noted that the mRMR was applied only to the training set in each fold, yet may be optimistic as the parameters were tuned during preliminary experiments.

Table 5-18 Depression Classification Performance using SVM and mRMR in 2-stage Feature Selection Approach of HOEIM-KN-ngram-Parameters (\( p_{11}, p_{12}, p_{21}, p_{22}, \theta_{11}, \theta_{12}, \theta_{21}, \theta_{22} \)) for Each Order (N) and Interaction

<table>
<thead>
<tr>
<th>Feature Set: mRMR 2-stage optimized HOEIM-KN-ngram parameters Classifier: SVM</th>
<th>EPI</th>
<th>FCI</th>
<th>PSI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Order (N)</td>
<td>Acc</td>
<td>Sens</td>
<td>Spec</td>
</tr>
<tr>
<td>1</td>
<td>81.0</td>
<td>72.4</td>
<td>88.2</td>
</tr>
<tr>
<td>2</td>
<td>74.6</td>
<td>69.0</td>
<td>79.4</td>
</tr>
<tr>
<td>3</td>
<td>82.1</td>
<td>79.3</td>
<td>85.2</td>
</tr>
<tr>
<td>4</td>
<td>93.7</td>
<td>93.1</td>
<td>94.1</td>
</tr>
<tr>
<td>5</td>
<td>77.8</td>
<td>79.3</td>
<td>76.5</td>
</tr>
</tbody>
</table>

Feature optimization was performed on HOEIM-KN-ngram parameters, not MMM, due to higher performance in EXP9. The SVM classification performance of the 2-stage mRMR optimized subset of HOEIM-KN-ngram parameters is given in Table 5-18. The 2\(^{nd}\), 3\(^{rd}\), 4\(^{th}\) and 5\(^{th}\) order were improved by 0.1%, 2.7%, 12.6% and 2.4% compared to the 1\(^{st}\) order. The best result was at the 4\(^{th}\) order at 93.7%, 93.8% and 98.9% for EPI, FCI and PSI. Consistently, the worst to best topics are EPI, FCI and PSI with an average accuracy across orders of 81.8%, 87.2% and 89.2%.
5.9.7 Comparison of HOEIM depression classification performance

The receiver operating characteristics (ROC), Figure 5-28, evaluate the sensitivity and specificity of all orders for three feature extraction methods (HOEIM-MMM, HOEIM-KN-ngram, mRMR HOEIM-KN-ngram) using an SVM.

For each interaction the performance of the sensitivity and specificity ratio from least to most sensible was HOEIM-MMM, HOEIM-KN-ngram and best with mRMR-optimized features. It was observed that PSI was the optimal interaction and EPI worst for all feature approaches.

![Figure 5-28 ROC space of SVM depression classification comparing performance of features extracted from three HOEIM variants.](image)

Table 5-19 supplies SVM accuracy, averaged over five orders, comparing HOEIM feature extraction approaches: HOEIM-MMM, HOEIM-KN-ngram and 2-stage mRMR HOEIM-KN-ngram features. The results showed HOEIM-MMM features performed lower than HOEIM-KN-ngram by an average of 6%, 4% and 3% for EPI, FCI and PSI.
Table 5-19 SUMMARY OF SVM DEPRESSION CLASSIFICATION ACCURACY, AVERAGED ACROSS HOEIM ORDERS, COMPARING HOEIM FEATURE EXTRACTION METHODS OF MMM AND KN-NGRAM INCLUDING WITH A 2-STAGE mRMR/SVM OPTIMIZATION

<table>
<thead>
<tr>
<th>Feature</th>
<th>EPI</th>
<th>FCI</th>
<th>PSI</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOEIM-KN-ngram mRMR</td>
<td>81.8</td>
<td>87.2</td>
<td>90.1</td>
</tr>
<tr>
<td>HOEIM-KN-ngram</td>
<td>66.8</td>
<td>79.2</td>
<td>78.5</td>
</tr>
<tr>
<td>HOEIM-MMM</td>
<td>60.9</td>
<td>75.3</td>
<td>75.5</td>
</tr>
</tbody>
</table>

It can be noted that every order and interaction HOEIM-KN-ngram, Table 5-16, was more accurate than HOEIM-MMM, Table 5-17. 2-stage mRMR optimization feature selection, Table 5-18, improved on the entire HOEIM-KN-ngram feature set, Table 5-16, by an average of 15%, 8% and 11.7%.

Another aspect is comparing accuracy of each HOEIM approach, given in Figure 5-29 averaged for topics. KN-ngram was more accurate than MMM approach by an average of 3%, 2%, 2%, 5%, and 9% for 1st to 5th order and optimized feature selection improved on the entire KN-ngram set by 13%, 8%, 9%, 20% and 9%.

![Figure 5-29](image_url)

Figure 5-29 Average depression classification accuracy, across the interactions (EPI, FCI, PSI), comparing the HOEIM feature extraction methods (HOEIM-MMM, HOEIM-KN-ngram and HOEIM-KN-ngram Optimized with mRMR) for each order
5.10 Discussion and Summary

A) Fit of various Influence Model (IM) implementations (DEIM and HOEIM)

Comparing EIM implementations found the DEIM, in terms of log-likelihood ($LL$) maximization, degraded as the delay increased. In contrast the $LL$ in HOEIM improved as the order increased and was worse for MMM compared to ngram and the best KN-smoothing. Thus the best model (i.e. maximum $LL$) was the 5th order HOEIM-KN-ngram.

B) Quantitative (Statistical) and Qualitative interoperations of EIM parameters

An examination of the proposed DEIM and HOEIM parameters (i.e. conditional probabilities and Influence Coefficients) delivered valuable insights towards discriminating adolescent depression risk factors. DEIM and HOEIM successfully expanded on the EIM by increasing the parameters an providing detailed conversation analysis as a trajectory of time delays (DEIM) or increased memory information (HOEIM) both describing how the emotional influences change over time.

The DEIM and HOEIM quantitative parameters (P and IC) were statistically analyzed using MANOVA and t-tests and found many features were statistically significant, $p<0.05$, different between depressed and non-depressed adolescents.

Additionally the parameters provided qualitative analysis with interoperations of emotional influence comparing depressed and control adolescents and agreed with psychological explanations of parent-child interactions. It is important to find and alter negative reciprocal parent-child processes and chronic maladaptive family-based interactions that reinforce depressive behaviors [137][135][144][146]. The EIM could offer a new tool to find depression associated risk factors of emotion regulation and possibly assist therapies to reduce maladaptive interactions to improve mental state.
C) Classification Performance using DEIM and HOEIM parameters

Utilizing features generated from the various EIMs (DEIM and HOEIM) towards automatic depression detection was achieved using classical machine learning models including an SVM and NN. In general the DEIM attained the best results with the larger delays and similar the HOEIM was better at higher order.

In the case of DEIM features a comparison of SVM and NN found the highest accuracy used NN and improved by optimizing the hidden nodes (O-NN). The O-NN using the original EIM (i.e. $t-1$) had as average accuracy for all topics of 87%. The DEIM extension improved with an average of 93% at $n=5$. Overall the best DEIM is for FCI/$t$-5 with 97.5% accuracy, 99.1% sensitivity and 93.3% specificity.

HOEIM classification was compared with different approaches (MMM and KN-ngram) and determined SVM depression classification performance was better using KN-ngram rather than MMM. An optimized subset, facilitated by a 2-stage mRMR method, improved performance of the entire feature set HOEIM-KN-ngram. The best classification setup, SVM/mRMR 4th order HOEIM-KN-ngram, reached an average accuracy of 95% across topics and a peak of 98.9% in PSI.

Figure 5-30 compares average accuracy of SVM adolescent depression classification using HOEIM and DEIM. The best results were achieved with the HOEIM compared to the DEIM with less feature parameters. On average HOEIM-MMM outperformed DEIM by 0%, 6%, 8%, 5% and 1% for the 1st to 5th order/delay. A greater improvement using HOEIM-KN-ngram with 3%, 8%, 10%, 10% and 9% increase for the respective order/delays. The mRMR feature optimization was the best overall SVM test, outperforming the DEIM by 16%, 16%, 19%, 30% and 18%.
D) Classification Performance Summary using DEIM and HOEIM parameters

Table 5-20 supplies a summarized ranking of depression detection performance based on conversation modeling parameters for each setup. This is a direct comparison of the original EIM, DEIM, HOEIM-MMM, HOEIM-KN-ngram and HOEIM-KN-ngram with mRMR optimized feature selection for every delay/order and classifier.

The EIM \((t-1)\) performed worse than the longer delays (DEIM) and more so the higher order (HOEIMs). A direct comparison of DEIM and HOEIM features for SVM depression classification found that the DEIM was obviously worse than the HOEIM approaches. Specifically, the HOEIM-MMM approach was less accurate than HOEIM-KN-ngram feature set. Overall it was optimal to use the 2-stage mRMR SVM feature selection method with a top average accuracy of 95.5% using the 4\(^{th}\) order HOEIM-KN-ngram mRMR SVM.
### Table 5-20: Ranking of Depression Classification Average Accuracy (%) of Each Interaction (EPI, FCI, PSI) Using the ORI-DB LIFE Annotations Construct States (positive, negative, neutral, silence) for the Influence Model Approach. The Cells Shaded Darkest to Lightest Denote the: EIM, DEIM, HOEIM-MMM, HOEIM-KN-Ngram and HOEIM-KN-Ngram using mRMR Feature Selection

<table>
<thead>
<tr>
<th>Accuracy (%)</th>
<th>Model Feature</th>
<th>Order/Delay</th>
<th>Classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>95.5</td>
<td>HOEIM-KN-ngram</td>
<td>4th</td>
<td>mRMR-SVM</td>
</tr>
<tr>
<td>93.4</td>
<td>DEIM</td>
<td>t-5</td>
<td>O-NN</td>
</tr>
<tr>
<td>92.3</td>
<td>DEIM</td>
<td>t-6</td>
<td>O-NN</td>
</tr>
<tr>
<td>90.8</td>
<td>DEIM</td>
<td>t-4</td>
<td>NN</td>
</tr>
<tr>
<td>90.8</td>
<td>DEIM</td>
<td>t-4</td>
<td>O-NN</td>
</tr>
<tr>
<td>89.6</td>
<td>DEIM</td>
<td>t-2</td>
<td>O-NN</td>
</tr>
<tr>
<td>88.8</td>
<td>DEIM</td>
<td>t-3</td>
<td>O-NN</td>
</tr>
<tr>
<td>86.6</td>
<td>DEIM</td>
<td>t-1</td>
<td>O-NN</td>
</tr>
<tr>
<td>85.5</td>
<td>HOEIM-KN-ngram</td>
<td>3rd</td>
<td>mRMR-SVM</td>
</tr>
<tr>
<td>85.2</td>
<td>HOEIM-KN-ngram</td>
<td>5th</td>
<td>mRMR-SVM</td>
</tr>
<tr>
<td>85.0</td>
<td>DEIM</td>
<td>t-8</td>
<td>O-NN</td>
</tr>
<tr>
<td>84.3</td>
<td>DEIM</td>
<td>t-5</td>
<td>NN</td>
</tr>
<tr>
<td>83.6</td>
<td>DEIM</td>
<td>t-3</td>
<td>NN</td>
</tr>
<tr>
<td>83.5</td>
<td>DEIM</td>
<td>t-10</td>
<td>O-NN</td>
</tr>
<tr>
<td>82.9</td>
<td>HOEIM-KN-ngram</td>
<td>2nd</td>
<td>mRMR-SVM</td>
</tr>
<tr>
<td>82.9</td>
<td>HOEIM-KN-ngram</td>
<td>1st</td>
<td>mRMR-SVM</td>
</tr>
<tr>
<td>82.2</td>
<td>DEIM</td>
<td>t-8</td>
<td>NN</td>
</tr>
<tr>
<td>81.8</td>
<td>DEIM</td>
<td>t-6</td>
<td>NN</td>
</tr>
<tr>
<td>80.5</td>
<td>DEIM</td>
<td>t-9</td>
<td>O-NN</td>
</tr>
<tr>
<td>79.4</td>
<td>DEIM</td>
<td>t-7</td>
<td>O-NN</td>
</tr>
<tr>
<td>79.3</td>
<td>DEIM</td>
<td>t-1</td>
<td>NN</td>
</tr>
<tr>
<td>78.8</td>
<td>DEIM</td>
<td>t-10</td>
<td>NN</td>
</tr>
<tr>
<td>78.3</td>
<td>DEIM</td>
<td>t-2</td>
<td>NN</td>
</tr>
<tr>
<td>76.8</td>
<td>HOEIM-KN-ngram</td>
<td>3rd</td>
<td>SVM</td>
</tr>
<tr>
<td>76.5</td>
<td>HOEIM-KN-ngram</td>
<td>5th</td>
<td>SVM</td>
</tr>
<tr>
<td>75.9</td>
<td>HOEIM-KN-ngram</td>
<td>4th</td>
<td>SVM</td>
</tr>
<tr>
<td>75.0</td>
<td>HOEIM-MMM</td>
<td>3rd</td>
<td>SVM</td>
</tr>
<tr>
<td>74.9</td>
<td>HOEIM-KN-ngram</td>
<td>2nd</td>
<td>SVM</td>
</tr>
<tr>
<td>72.8</td>
<td>DEIM</td>
<td>t-7</td>
<td>NN</td>
</tr>
<tr>
<td>72.7</td>
<td>HOEIM-MMM</td>
<td>2nd</td>
<td>SVM</td>
</tr>
<tr>
<td>70.6</td>
<td>HOEIM-MMM</td>
<td>4th</td>
<td>SVM</td>
</tr>
<tr>
<td>70.0</td>
<td>HOEIM-KN-ngram</td>
<td>1st</td>
<td>SVM</td>
</tr>
<tr>
<td>69.7</td>
<td>DEIM</td>
<td>t-9</td>
<td>NN</td>
</tr>
<tr>
<td>67.6</td>
<td>HOEIM-MMM</td>
<td>5th</td>
<td>SVM</td>
</tr>
<tr>
<td>67.1</td>
<td>DEIM</td>
<td>t-1</td>
<td>SVM</td>
</tr>
<tr>
<td>67.1</td>
<td>HOEIM-MMM</td>
<td>1st</td>
<td>SVM</td>
</tr>
<tr>
<td>67.1</td>
<td>DEIM</td>
<td>t-5</td>
<td>SVM</td>
</tr>
<tr>
<td>67.1</td>
<td>DEIM</td>
<td>t-2</td>
<td>SVM</td>
</tr>
<tr>
<td>66.7</td>
<td>DEIM</td>
<td>t-3</td>
<td>SVM</td>
</tr>
<tr>
<td>65.7</td>
<td>DEIM</td>
<td>t-4</td>
<td>SVM</td>
</tr>
<tr>
<td>65.6</td>
<td>DEIM</td>
<td>t-7</td>
<td>SVM</td>
</tr>
<tr>
<td>65.4</td>
<td>DEIM</td>
<td>t-9</td>
<td>SVM</td>
</tr>
<tr>
<td>65.3</td>
<td>DEIM</td>
<td>t-6</td>
<td>SVM</td>
</tr>
<tr>
<td>65.1</td>
<td>DEIM</td>
<td>t-8</td>
<td>SVM</td>
</tr>
<tr>
<td>62.7</td>
<td>DEIM</td>
<td>t-10</td>
<td>SVM</td>
</tr>
</tbody>
</table>
Chapter Six:

PREDICTION OF EMOTIONAL INTERACTIONS IN CONVERSATIONS

6.1 Preview
Prediction of emotions plays a particularly important role in generation of naturalistic human-machine conversations. Understanding conversational interaction patterns is an important concept for social science, behavioral analysis and psychological purposes. This chapter explores adolescent affect prediction from knowledge of past constructs (positive, negative, neutral and silence) from the labeled parent-adolescent conversations previously described in Section 5.2 with the ORI-DB from Section 3.3.

Methods investigated include random walks (RW) and a new interacting RW (IRW) both outlined in Section 6.2.1. Another approach of non-linear autoregressive models (NAR/NARX) implemented with NN and ANFIS as described in Section 6.2.2. The experimental setup for RW and NAR experiments are described in Section 6.3, results are supplied and evaluated in Section 6.4 and a conclusion in Section 6.5.

6.2 Methodology

6.2.1 Introduction of the Random Walk
A random walk is a mathematical formalization of a path consisting of random step first introduced in 1905 [51]. Considering a particle that can only move one step with equal probability, the RW can explain the observed behaviors of many processes.
6.2.1.1 Biased Random Walk

The simple random walk (RW) is based on the concept that a particle moves with equal probability in any direction. A RW can be biased leading to a net drift in a specific direction, known as a biased RW, using either of the following mechanisms:

- Instead of equal probability of moving either direction there is a bias towards higher probability of moving in one direction over the others
- Probability of moving in each direction is equal with non-constant distance.

The RW is considered as a Markov chain with a finite state space $S = \{s_1, \ldots, s_n\}$, initial probability vector $\pi_0 = (p_1, p_2, \ldots, p_n)$, and transition probabilities matrix $P = (p_{ij})$. The RW produces a sequence, $X_0, X_{t-1}, X_{t-2}, \ldots$, such that $X_0$ is chosen according to the probability vector $\pi_0$ and each new $X_{t+1}$ according to the transition probability matrix and $X_t$. A simulation of a possible process is as follows:

1. The system has a finite number of states with the state space as follows:

   $$S = \{s_1, s_2, \ldots, s_n\}$$

   In our case the state-space consists of conversational descriptors:

   $$S=\{\text{positive, negative, neutral, silence}\}$$

2. Starting in a random initial state at $t = 0$, the system changes state randomly at $t = 1, 2, \ldots$, with the evolution described by a chain of random variables

   $$X_0 \rightarrow X_1 \rightarrow X_2 \rightarrow \cdots X_t \in S$$

3. At $t = 0$, the system occupies state $s_k$ with initial probability $\pi_0 = (p_1, p_2, \ldots, p_n)$ such that:

   $$\pi_0(s_k) = P(X_0 = s_k), k = 1, 2, \ldots$$

4. Suppose that at any time the system has the following history: $X_i = x_i, i = 0, 1, 2, \ldots$. At a particular time the system the next step only depends on the present, given by the transition probability from state $s_i$ to $s_j$ as follows:

   $$p_{ij}(t) = P(X_{t+1} = s_j | X_t = s_i).$$
6.2.1.2 Extension to higher order random walk (HORW)

Prediction of a future emotional (construct) state could be improved based on the current state and furthermore previous state(s). This was attained with higher order Markov transition probabilities, estimated with KN-N-grams that concatenates previous states into a meta-state as a 1st order representation of higher orders.

6.2.1.3 New interacting random walk (IRW)

The series, $X_0, X_1, \ldots, X_t \in S$, created from the RW described above was based on self-transition probabilities so the next state was predicted based on that sequence’s current and past state(s) with no external factors. This was more realistic in conversations that external factors, such as other speakers’ affect emotional state.

It was proposed to extend the RW into a new interacting RW by including another speakers sequence as an external series. The self and cross-transition probabilities were integrated so the next state of speaker $i$ was based on previous states of $i$ and $j$ and weighted by Influence Coefficients, from Chapter Five, to reflect the relative importance of each speaker.

6.2.1.4 2D representation of the random walk methods

The 1D RW can be extended to represent a Markov chain along a plane (2D) or space (3D). In this application the speaker was visualized as a RW on a directed 2D graph, $G(V,E)$, with a set of vertices, $V$, connecting edges, $E$, and moved in proportion to the probability matrix [117]. Each direction was mapped to the state space $S$, shown in Figure 6-1, positive moves up, negative down, silence left and neutral right.
Figure 6-1 2D random walk on a graph illustrating the direction the speaker (particle) can move into either the positive (up (+)), negative (down (-)), neutral (right (N)) or silent (left (S)) state. The example demonstrates a speaker that starts at the origin (0,0) denoted as position u and then moves up (0,1) to position v into a positive state (+) with a probability of \( p_{uv} \) and then moves right (1,1) to position w in a neutral state (N) with a probability of \( p_{vw} \).

Given an initial position at vertex u a move to a neighboring vertex v is determined by the transition probability \( P(u,v) \) and a random perturbation to ensure less probable states are travelled. For a graph with probability matrix \( P = (p_{uv})_{u,v \in V} \) the transition probability of u to v is non-zero if \( (u,v) \in E \). The model converges to a stationary probability distribution associated with graph vertices [117].

Figure 6-2 2D Random Walk process from the origin (0,0) to the final position \((x_f, y_f)\) with phasor notation to represent the magnitude of distance travelled, \( |D| \), and the drift being defined by the angle, \( \theta \), travelled.
The RW was visualized as drifting towards an overall affect defining emotional drift. If a walk tended upwards it was generally positive, below mostly negative and left and right for turn-taking analysis and dominance; useful for psychological analysis.

Instead of a sequence, \( X_t, X_{t-1}, X_{t-2}, \ldots \), belonging to state space \( S \) it is converted to a 2D graph in the \( x \) and \( y \) position. The new sequence \( Z_t, Z_{t-1}, Z_{t-2}, \ldots \), represents the walker on the graph shown by Figure 6-2. The position is represented in phasor notation in the complex plane after \( N \) steps by equation (6.1) \cite{118}.

\[
z = \sum_{j=1}^{N} e^{i\theta_j}
\]

(6.1)

The coordinates \( (x_f, y_f) \) at speaker’s last step represent the final position defined by the last state in the sequence \( Z_t \). Considering this is a simple lattice the final position was found by summing the random walk sequence in the \( x \) and \( y \) axis with plus/minus one unit per step given by equations (6.2) and (6.3).

\[
x_f = \sum_{j=1}^{N} x_j
\]

(6.2)

\[
y_f = \sum_{j=1}^{N} y_j
\]

(6.3)

The final position of the random walk was measured in RMS distance, to normalize against the number of steps, given by equation (6.4) and the angle by (6.5).

\[
|D_{\text{RMS}}| = \sqrt{\frac{x_f^2 - y_f^2}{N}}
\]

(6.4)

\[
\theta = \tan^{-1}\left(\frac{y_f}{x_f}\right)
\]

(6.5)
6.2.2 Non Linear Autoregression (NAR) models for prediction

Another time-series prediction method using uses autoregression (AR) models via linear regression, given by equation (6.6), where current value $y_t$ is based on a weighted linear sum of past states, $y_{t-p}$, where $\varepsilon_t$ and $b$ are noise and constant values.

$$y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + \varphi_p y_{t-p} + \varepsilon + b \quad (6.6)$$

Real-life systems generally cannot be fully explained with a linear AR model and require non-linear models [41][252][253]. The current value of the time series is related to past values as a non-linear function $f$, shown by Figure 6-3, such as a polynomial, neural network, wavelet network or sigmoid network.

![Figure 6-3 General process of a Non-linear Autoregressive model (NAR) such that the output, $y(t)$, is a function of its own past states, $y(t-1), \ldots, y(t-n)$](image)

The NAR model can predict emotional states, $y(t)$, of a child based on their historical data, $y(t-n)$. However, when the children are engaged in discussions with their parents there is an external influence to emotional states. It is useful to incorporate the parents as practical predictors as an external (exogenous) input.

To achieve this a nonlinear autoregressive with an exogenous input (NARX) is used, as shown by the network in Figure 6-4. Now, $y(t)$ is predicted using both $y(t)$ and an exogenous input $u(t)$ respectively the child and parent affect states.

![Figure 6-4 General process of a Non-linear Autoregressive model with external input (NARX) describing the output, $y(t)$, as a function of its own past states, $y(t-1), \ldots, y(t-n)$, and the past states of the external input, $u(t-1), \ldots, u(t-m)$](image)
6.2.2.1 NAR with Neural Networks (NAR-NN and NARX-NN)

NN are popular for time series analysis and NAR/NARX implementation and have been used in prediction applications for decades [76]. NN can approximate any non-linear function given enough training data and hidden nodes [76] and are considered a universal estimators [315][316]. To learn time varying patterns for time-series prediction multiple past time delays are used in dynamic networks. The dynamic network memory makes it more powerful than a static network [343].

The dynamic network memory can be either with feed forward connections using delays only at the model input or a recurrent network incorporating memory using loops within intermediate layers. The recurrent network is more difficult to train and slower than the feed forward network [148]. The recurrent NN and NAR-NN have been compared and there is no theoretical proof there is an advantage looping back layer outputs and minimal improvement with the recurrent NN [148].

For these reasons it is understandable that the dynamic network not be recurrent and only use memory at the input. Applying a tapped time-delay line (TDL) at the Neural Network input produces the dynamic network. The time-delay Neural Network (TDNN) is capable of learning from the multiple past instances.

![NAR Neural Network Architecture](image)

**Figure 6-5** NAR Neural Network architecture, adapted from [254], illustrating estimated output, $\hat{y}(t)$, is based on the time-delay line (TDL), the corresponding weights, $w$, and bias, $b$, parameters of the NN along with hidden layer and output layer activation functions, $f$. 

201
The non-linear autoregressive (NAR) model can be implemented using the NN with a TDL as memory. The NAR-NN is shown in Figure 6-5, where the TDL is given at the input as a vector \( y(t) \), corresponding to the child’s past sequence. Similarly, using NAR with an exogenous input, the NARX-NN is illustrated by Figure 6-6 with two TDL signifying the external input \( u(t) \) (parent) and the child’s past states \( y(t) \).

![Figure 6-6 NARX-NN architecture, adapted from [254], illustrating estimated output, \( \hat{y}(t) \), based on time-delay line (TDL) of the output and external input, \( u(t) \), with corresponding weights, \( w \), and bias, \( b \), parameters and activation functions, \( f \).](image)

The networks apply tapped delay lines (TDL) to store previous values of speaker sequences. The architecture of the network was a two-layer feed forward NN with one hidden layer with 100 nodes. A tan-sigmoid transfer function in the hidden layer \( (f1) \) and a linear transfer function \( (f2) \) in the output layer; shown in Figure 6-7.

![Figure 6-7 Activation functions, \( f \), used for the hidden and output layers illustrating a tan-sigmoid (left) and a linear activation function (right).](image)
The output of the network is the estimated value of affect, \( \hat{y}(t) \), which can be fed back to the input as a past state, shown in Figure 6-8, as a closed loop system. The performance is improved using the true output in open loop architecture, which is possible if the outputs are known. Another advantage of an open loop is that it is purely feed forward so a static back propagation algorithm is efficient [115][79].

Figure 6-8 Feed forward NARX architectures used to estimated a sequence, \( \hat{y}(t) \), with a tapped time delay line (TDL) of past states of an external input, \( u(t) \), and past states of the estimated output, \( \hat{y}(t) \), fed back in a closed loop (left) or using actual output, \( y(t) \), in an open loop (right) replicated from [254]

To learn the weight and bias values of the network the training was performed by minimizing the non-linear least squares, given by equation (6.7), from the MATLAB toolbox provided in [161]. Least squares was solved using Levenberg-Marquardt back-propagation algorithm, which interpolates between the Gauss-Newton algorithm and the gradient decent method [159][160].

\[
\min_x \| F(x) \|^2 = \min_x \sum_i f_i^2(x) + f_2^2(x) + \cdots + f_n^2(x) \quad (6.7)
\]
6.2.2.2 NAR with Adaptive Neuro-Fuzzy Inference System (ANFIS)
Additionally, NAR and NARX were implemented using the Adaptive Neuro-Fuzzy Inference System (ANFIS) denoted as NAR-ANFIS and NARX-ANFIS. ANFIS integrates a fuzzy interference system (FIS), corresponding to a set of fuzzy if-then rules, capable of learning nonlinear functions and is a universal estimator [110][111].

For a set of input/output data sets the ANFIS constructs a FIS with input and output membership functions and associated parameters. The ANFIS program offers three options to initialize the Sugeno-type FIS structure before ANFIS fine-tuning:

- grid-partition (GP)
- Subtractive clustering (SC)
- fuzzy c-means clustering (FCM)

Grid partitioning generates fuzzy rules by enumerating all combinations of membership functions, $M$, of all inputs, $N$. This issue is the ‘curse of dimensionality’ as the number of rules increases exponentially with the input size, $M^N$. It is suggested FIS is generated using GP with less than 6 inputs otherwise SC [41][44][110].

After FIS is initialized ANFIS uses an optimization algorithm to adjust the parameters with fine-tuning using a hybrid of back propagation gradient decent in combination with least squares estimation [44]. ANFIS creates a fuzzy decision tree into $p^n$ linear regression models to minimize the sum of the squared errors:

$$SSE = \sum_j e_j^2$$  \hspace{1cm} (6.8)

$e_j$ is the error between actual output and desired

$p$ is the number of fuzzy partitions of each variable (membership functions)

$n$ is the number of input variables
The NAR-ANFIS input layer has one input vector, shown by Figure 6-9, representing the modeled time series, \( y(t) \). NARX has two inputs both \( y(t) \) and an external input \( u(t) \) shown by Figure 6-10. ANFIS is a dynamic network with memory that can learn time-varying patterns from previous speaker states \((u(t) \text{ and } y(t))\).

Figure 6-9 NAR Network implemented with ANFIS using past states of \( y \) at the input to the \( K \) membership functions \((A_k)\) to generate the estimated output, \( \hat{y} \), as the normalized weighted, \( \hat{\omega}_k \), summation of each of the membership function outputs, \( f_k \).

Figure 6-10 NARX implemented with ANFIS using past data of \( y \) and external sequence \( u \) at the \( K \) input membership functions \((A_k \text{ and } B_k)\) to estimate the next state, \( \hat{y} \), as the normalized weighted, \( \hat{\omega}_k \), summation of each output MF result, \( f_k \).
6.3 Experimental Setups

The experiments in this section predicted the next affect state of a speaker based on historical labeled data from the ORI-DB described in Chapter Three and the annotation corpus from Section 5.2, of four constructs (positive, negative, neutral and silence). Prediction was explored from two main approaches: Random Walks and Non-Linear Autoregressive Models.

6.3.1 Experimental Setup Random Walks

The random walk simulated possible child affect sequences in conversations with parents, as described in Section 6.2.1. The performance is measured as the average relative error between the actual probability matrix ($P_{\text{ref}}$) and the final converged matrix of the RW ($P_{\text{rw}}$) and the accuracy defined by equations (6.9) and (6.10).

\[
\text{Error} = \frac{P_{\text{rw}} - P_{\text{ref}}}{P_{\text{ref}}} \quad (6.9)
\]

\[
\text{Accuracy} = (1 - \text{Error}) \times 100\% \quad (6.10)
\]

The experiments of Random walks are as follows:

- **EXP1**: Determined importance of memory in the RW for affect prediction

- **EXP2**: Investigated the use of an interacting random walk (IRW) to predict children’s affect based on their own and their parents past affect states.

- **EXP3**: 2D random walks were examined in relation to emotional drift and affect prediction in depressed and control adolescents
6.3.2 Experimental Setup NN and ANFIS methods of NAR(X) models

The NN and ANFIS implemented NAR/NARX for adolescent affect prediction, follows Section 6.2.2, using early stopping cross validation to avoid over fitting. The annotated data is separated into 70% training, 15% validation and 15% test sets with 5-fold CV. ANFIS implementation is associated with 3 Gaussian curve input membership functions. The NN implementation used a feed forward NN with a hidden layer of 100 nodes, a tan-sigmoid transfer function in the hidden layer and a linear transfer function in the output layer. The experiments are briefly explained below with more detail in the results section.

- **EXP1**: Determined the ideal memory required for predicting future affects in conversations using time delay line in an autoregressive Neural Network using speaker and class independence (SI-CI)

- **EXP2**: Investigated adolescent affect prediction with the inclusion of the using parents affect as an eternal input to an exogenous NN (NARX) (SI-CI)

- **EXP3**: Repeated EXP1 and EXP2 and determined if depression is a factor for affect prediction by including class dependence (SI-CD) NAR and NARX

- **EXP4**: Investigated if speaker dependence can improve the performance of affect prediction by repeating EXP1 and EXP2 (SD-CI)

- **EXP5**: Similarly experimented using class dependence and speaker dependence using the setup from EXP1 and EXP2 (SD-CD)

- **EXP6**: Used the optimal setup speaker and class dependent setup (SD-CD), as determined in the NN experiments (EXP1-EXP5), and investigated ANFIS implemented NAR/NARX affect prediction.
The following experimental cases for the prediction of has been tested for the Neural Network autoregressive models both the NAR and NARX up the 5th order:

(1) Speaker independent and class independent (SI-CI) case, data representing all children and parents from both class was train together with NAR-NN and NARX-NN models. (*EXP1* and *EXP2*)

(2) Speaker independent and class dependent (SI-CD) case with two sub-cases:
- SI-CD for depressed adolescents (SI-CD-D)
- SI-CD for non-depressed adolescents (SI-CD-ND)

The speaker independent and class dependent (SI-CD) case, data representing all children that belonged to a given class was used to train the models separately for the depressed and control classes. (*EXP3*)

(3) Speaker dependent and class independent (SD-CI) modeling, data representing each child of a given and the corresponding parent was trained to predict the child's sequence of states. (*EXP4*)

(4) Speaker dependent and class dependent (SD-CD) case with two sub-cases:
- SD-CD for depressed adolescents (SD-CD-D)
- SD-CD for non-depressed adolescents (SD-CD-ND)

In the speaker dependent and class dependent (SD-CD) modeling, data representing each child of a given class (SD-CD-D or SD-CD-ND) and the corresponding parent was trained separately to predict the child's affect (*EXP5*).
6.4 Experimental results

6.4.1 Experimental results implemented with RW

6.4.1.1 Effectiveness of higher order random walks (HORW) (EXP1)

Table 6-1 gives the average relative adolescent affect prediction accuracy of four ORI-DB constructs generated by the HORW using self-transition \( P_{ii} \), outlined in Section 6.2.1.2. The results showed an improvement as the order of the system increased. This would be intuitive, as more history/memory of states would help determine the next state.

<table>
<thead>
<tr>
<th>Order</th>
<th>Iteration/steps</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10</td>
</tr>
<tr>
<td>1</td>
<td>78.0</td>
</tr>
<tr>
<td>2</td>
<td>80.9</td>
</tr>
<tr>
<td>3</td>
<td>87.6</td>
</tr>
<tr>
<td>4</td>
<td>92.7</td>
</tr>
<tr>
<td>5</td>
<td>96.9</td>
</tr>
</tbody>
</table>

6.4.1.2 Effectiveness of the interacting random walk (IRW) (EXP2)

The average accuracy of the interacting random walk, outlined in Section 6.2.1.3, using IC and both \( P_{ii} \) and \( P_{ij} \), is given in Table 6-2. The results were consistent with the previous model in that the IRW improved with order and increase in steps.

<table>
<thead>
<tr>
<th>Order</th>
<th>Iteration/steps</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10</td>
</tr>
<tr>
<td>1</td>
<td>82.0</td>
</tr>
<tr>
<td>2</td>
<td>83.7</td>
</tr>
<tr>
<td>3</td>
<td>90.0</td>
</tr>
<tr>
<td>4</td>
<td>94.8</td>
</tr>
<tr>
<td>5</td>
<td>97.9</td>
</tr>
</tbody>
</table>
The average relative error of the IRW and HORW are compared in Figure 6-11 showing the improvement of IRW reduced as the number of steps grew (i.e. 2.5%, 0.9%, 0.04% and -1.3%). Moreover as the order increased the performance difference between IRW and HORW diminished from 1.3%, 0.5%, 0.27%, 0.43% and 0.06%, suggesting the addition of ICs and external parent influence decreased in importance as the time lag increased.

![Figure 6-11 Average relative error (%) of the Higher Order Random Walk (HORW) (filled bar) in comparison to the Interacting Random Walk (IRW) (hashed bar) for orders 1-5 and each iteration length (10, 25, 50 and 250 steps)](image)

6.4.1.3 Investigate 2D RW for link between depressed interaction (EXP3)

The 2D random walk, outlined in Section 6.2.1.4, was evaluated by comparing the estimated RW and the actual path taken by the speaker. The performance was based on the relative error of the RMS distance, $|D_{RMS}|$, and the angle, $\theta$, between the final positions of actual and estimated walks.

![Figure 6-12 Average relative difference of distance $|D_{RMS}|$ and angle $\theta$ between real data and the random walks of each order for 25 (left) and 250 steps (right)](image)
The results are given in Figure 6-12; comparing the 2D RW and the IRW. Both measures (|DRMS| and $\theta$) reduced as the order increased until the 3rd or 4th order then degraded. The angle was generally slightly closer to the raw data for the IRW compared to the RW. However, the |DRMS| always had a lower relative error than the RW.

The 2D random walks were biased towards a particular quadrant, +N, +S, -N and -S, as shown in Figure 6-13, and can be used to characterize emotional drift. Each quadrant corresponds to information regarding the leading sign (positive or negative) of the affect and the dominance based on silence.

![Figure 6-13 Representation of a random walk on a 2D graph of a speaker with positive, negative, neutral, silence in each direction with defined quadrants that characterize emotional drift. Each quadrant corresponds to information regarding the leading sign (positive or negative) of the affect and the dominance based on silence.](image)

The 2D random walks were biased towards a particular quadrant, +N, +S, -N and -S, as shown in Figure 6-13, and can be used to characterize emotional drift. The results in Table 6-3 show the accuracy of RW ending up in the correct quadrant. This accuracy improved with order and steps for both methods RW and the IRW.

<table>
<thead>
<tr>
<th>Order</th>
<th>25 iterations</th>
<th>250 iterations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$RW$</td>
<td>$IRW$</td>
</tr>
<tr>
<td>1</td>
<td>67.7</td>
<td>75.4</td>
</tr>
<tr>
<td>2</td>
<td>69.7</td>
<td>77.3</td>
</tr>
<tr>
<td>3</td>
<td>71.4</td>
<td>78.5</td>
</tr>
<tr>
<td>4</td>
<td>73.8</td>
<td>76.2</td>
</tr>
<tr>
<td>5</td>
<td>71.9</td>
<td>75.6</td>
</tr>
</tbody>
</table>

Table 6-3 Average Accuracy (%) of Correct Final Quadrant Position Comparing the Interacting Random Walk (IRW) and Random Walk (RW) for Two Different Iteration Lengths (25 and 250 Steps) and Orders 1-5
The 2D random walk was visualized on a graph, defined by Figure 6-13, depicting a drift towards positive (up y-axis), negative (down the y-axis), neutral (right x-axis) and silence (left x-axis). A few examples are given in Figure 6-14, where the blue lines are the actual speaker compared to the estimated paths and generally follow the same trend.

![Figure 6-14](image)

Figure 6-14 Three random walks for 250 steps comparing to the actual data in blue and another three examples (RW path 1, path 2 and path 3) showing different emotional drift in the x and y direction.

The 2D RW visualized the path of a speaker during an interaction such that the path corresponds to the emotional constructs taken at each step. Three examples of sequences with different trends were simulated by RWs given in Figure 6-15.

![Figure 6-15](image)

Figure 6-15 250-step random walks showing three different emotional drifts, negative-silent (green), positive-silent (black), positive-neutral (red).
Turn taking was analyzed such that the green and black speakers tended to the negative side of the x-axis that corresponds to more time in the silent state. In contrast it was observed that red walker spoke for the majority of the time, reflecting conversation dominance. Additionally, the side of the y-axis describe emotional valance such that both the black and red were more positive and green more negative.

![2D Random Walk](image)

Figure 6-16 2D graph representation of the RWs comparing the emotional drift of the 29 depressed (left) and 34 non-depressed (right) adolescents

A group of depressed children’s emotional steps were simulated with a RW and shown in Figure 6-16 for depressed (left) adolescents and non-depressed (right). The distinction for depressed adolescents was observed as a tendency to drift towards negative construct compared to the non-depressed, which is mostly positive. This agreed with psychology literature that states depression consists of emotional disturbances and prolonged phases of excessive sadness [449].

The 2D RW results presented were intended to show the potential and benefits of tracking emotional transition patterns and emotional drifts a speaker. In this application it has been shown that a subject’s depressive state could be visualized with undesirable interactions (e.g. trend towards negative states in depressed subjects). Medical practitioners could use the 2D RW patterns to test the effectiveness of treatments and help guide behavioral therapies in parent-adolescent interactions.
6.4.2 Experimental results implemented with NAR models

6.4.2.1 Prediction using Speaker and Class Independent (SI-CI) NAR-NN (EXP1)
NAR-NN was trained and tested for speaker and class independent (SI-CI) affect prediction, of four constructs, using the method outlined in Section 6.2.2.1. SI-CI NAR-NN affect prediction accuracy is given in Table 6-4 for each order and topic. In every setup the results improved with order possibly due to increased memory.

Table 6-4 AVERAGE ACCURACY (%) OF AFFECT PREDICTION USING THE SPEAKER INDEPENDENT AND CLASS INDEPENDENT (SI-CI) NAR-NN WITH A COMPARISON OF ORDERS FOR EACH INTERACTION (EPI, FCI, PSI). BOLD CELL INDICATES THE INTERACTION AND ORDER WITH THE HIGHEST ACCURACY FOR THE GIVEN SETUP

<table>
<thead>
<tr>
<th>Order</th>
<th>SI-CI NAR-NN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EPI</td>
</tr>
<tr>
<td>1</td>
<td>46.3</td>
</tr>
<tr>
<td>2</td>
<td>50.2</td>
</tr>
<tr>
<td>3</td>
<td>60.6</td>
</tr>
<tr>
<td>4</td>
<td>66.9</td>
</tr>
<tr>
<td>5</td>
<td>76.5</td>
</tr>
</tbody>
</table>

6.4.2.2 Effect of external factor (NARX-NN) on prediction (SI-CI) (EXP2)
An extension incorporated the parent’s affect states as an external input and generated a NARX-NN. The speakers were weighted by NN parameters instead of IC as used in the IRW. The NARX-NN SI-CI setup performance is given in Table 6-5 and showed accuracy increased with increase in order (memory).

Table 6-5 AVERAGE ACCURACY (%) OF AFFECT PREDICTION USING THE SPEAKER INDEPENDENT AND CLASS INDEPENDENT (SI-CI) NARX-NN WITH A COMPARISON OF ORDERS FOR EACH INTERACTION (EPI, FCI, PSI)

<table>
<thead>
<tr>
<th>Order</th>
<th>SI-CI NARX-NN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EPI</td>
</tr>
<tr>
<td>1</td>
<td>52.0</td>
</tr>
<tr>
<td>2</td>
<td>57.7</td>
</tr>
<tr>
<td>3</td>
<td>70.0</td>
</tr>
<tr>
<td>4</td>
<td>78.0</td>
</tr>
<tr>
<td>5</td>
<td><strong>89.6</strong></td>
</tr>
</tbody>
</table>

A direct comparison of NAR and NARX is given in Figure 6-17 as an average accuracy for all topics. NARX was on average 6.7%, 7.6%, 9.0%, 10.8% and 8.4% more accurate for orders 1-5. This suggested prediction based on coupled emotional interactions of two speakers improved prediction compared to an individual speaker.
6.4.2.3 Determine NAR(X)-NN depression dependence (SI-CD) (EXP3)

Continuing with speaker independent modeling the previous experiments were repeated with a class dependent model (SI-CD). This was achieved by creating separate models for the depressed and non-depressed groups. The results, given in Table 6-6, were consistent with previous results in relation to the advantage of higher orders and improvement with the inclusion of the external input (NARX).

Table 6-6 COMPARISON OF NAR-NN AND NARX-NN AVERAGE AFFECT PREDICTION ACCURACY (%) USING SPEAKER INDEPENDENT-CLASS DEPENDENT MODEL FOR DEPRESSED (SI-CD-D) (LEFT) AND NON-DEPRESSED (SI-CD-ND) (RIGHT) MODELS

<table>
<thead>
<tr>
<th>Order</th>
<th>SI-CD-D</th>
<th>SI-CD-ND</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NAR-NN</td>
<td>NARX-NN</td>
</tr>
<tr>
<td></td>
<td>EPI</td>
<td>FCI</td>
</tr>
<tr>
<td>1</td>
<td>51.6</td>
<td>53.6</td>
</tr>
<tr>
<td>2</td>
<td>58.9</td>
<td>60.6</td>
</tr>
<tr>
<td>3</td>
<td>68.1</td>
<td>69.7</td>
</tr>
<tr>
<td>4</td>
<td>82.2</td>
<td>79.7</td>
</tr>
<tr>
<td>5</td>
<td>88.2</td>
<td>82.2</td>
</tr>
</tbody>
</table>

The results showed that the NARX-NN was better than the NAR-NN in each case (order, SI-CD setup and interaction). The NARX-NN improved on the NAR-NN by an average of 6%, 9%, 8%, 5% and 4% for each order in the SI-CD-D setup. Similarly, the NARX-NN is 8%, 11%, 4%, 4% and 8% better than NAR-NN in the SI-CD-N setup.
The speaker-independent and class-dependent (SI-CD) predictions were more accurate than the speaker-independent class-independent (SI-CI) setups. On average for the NAR-NN case the S-CD-D (depressed) and SI-CD-ND (non-depressed) were 7.2% and 6.9% more accurate compared to the SI-CI. Similar, SI-CD-D and SI-CD-ND were 5.3% and 5.4% more accurate than the SI-CI setup for NAR-NN. This suggested that class dependence is an important factor for efficient affect prediction.

6.4.2.4 Compare NAR(X)-NN with Speaker Dependence (SD-CI) (EXP4)

Generally it was easier to obtain higher recognition rates from a speaker-dependent than a speaker-independent system [60][177][258][176]. Most studies show speaker dependence in advantageous but most work concentrates independence [233][234].

The average results, for all 63 conversational pairs, are given in Table 6-7 for speaker dependent and class independent (SD-CI) NAR and NARX NN. The NARX-NN was on average, for all topics and participants, 4.8%, 8.6%, 7.2%, 4.7% and 6.1% more accurate than the NAR for order 1-5 respectively (average of 6.3%).
Table 6-7 COMPARISON OF NAR-NN AND NARX-NN AVERAGE AFFECT PREDICTION ACCURACY (%) FOR CLASS INDEPENDENT AND SPEAKER DEPENDENT MODEL (SD-CI)

<table>
<thead>
<tr>
<th>Order</th>
<th>SD-CI</th>
<th>NAR-NN</th>
<th>NARX-NN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>EPI</td>
<td>FCI</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>57.4</td>
<td>61.8</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>60.4</td>
<td>66.1</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>71.8</td>
<td>70.0</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>81.3</td>
<td>81.6</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>89.0</td>
<td>86.9</td>
</tr>
</tbody>
</table>

Comparing the speaker independent model (SI-CI), from Sections 6.4.2.1 and 6.4.2.2, with this speaker dependent model (SD-CI) revealed that SD-CI was better than SI-CI for both NAR-NN and NARX-NN setups. A paired T-test, between SD-CI and SI-CI, shows statistically significant improvement for SD-CI (NAR \( p < 0.0001 \) and NARX \( p < 0.0009 \)), agreeing with literature [60][177][258][176].

![Average affect prediction accuracy (%) across topics (EPI, FCI, PSI), of class independent model (CI) comparing speaker independence (SD) and speaker dependence (SD) models for both NARX-NN NAR-NN models](image)

Specifically, the NAR-NN using SD-CI was on average 12.2%, 11.7%, 14.0% and 9.5% more accurate for each order compared to SI-CI, as shown by Figure 6-19. Similarly, the NARX-NN was optimal for the SD-CI model compared to SI-CI by 11.3%, 12.7%, 9.0%, 7.8% and 7.3%. The total average improvement using the SI-CI was 11.6% for the NAR and 9.6% for the NARX compared to speaker independence.
6.4.2.5 Compare NAR(X)-NN with Speaker and Class Dependence (SD-CD) (EXP5)

The average accuracy of speaker and class dependent (SD-CD) NAR and NARX are given for depressed (left) and non-depressed (right) in Table 6-8. In can be noted this is the weighted accuracy, based on group sizes, of SD-CI in Table 6-7. Overall the average accuracy of NAR-NN and NARX-NN are 2.4% and 2.1% higher in SD-CD-D compared to SD-CD-ND implying affect prediction is easier in depressed subjects.

**Table 6-8 COMPARISON OF NAR-NN AND NARX-NN AVERAGE AFFECT PREDICTION ACCURACY (%) USING SPEAKER DEPENDENT-CLASS DEPENDENCE MODEL WITH 29 DEPRESSED CONVERSATIONS (SD-CD-D) (LEFT) AND 34 NON-DEPRESSED CONVERSATIONS (SD-CD-ND) (RIGHT)**

<table>
<thead>
<tr>
<th>Order</th>
<th>SD-CD-D</th>
<th></th>
<th></th>
<th>SD-CD-D</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NAR-NN</td>
<td>NARX-NN</td>
<td></td>
<td>NAR-NN</td>
<td>NARX-NN</td>
<td></td>
</tr>
<tr>
<td></td>
<td>EPI</td>
<td>FCI</td>
<td>PSI</td>
<td>EPI</td>
<td>FCI</td>
<td>PSI</td>
</tr>
<tr>
<td>1</td>
<td>61.9</td>
<td>58.4</td>
<td>66.1</td>
<td>66.3</td>
<td>68.4</td>
<td>70.8</td>
</tr>
<tr>
<td>2</td>
<td>62.8</td>
<td>63.4</td>
<td>66.7</td>
<td>70.2</td>
<td>74.2</td>
<td>77.7</td>
</tr>
<tr>
<td>3</td>
<td>74.5</td>
<td>68.6</td>
<td>73.9</td>
<td>78.0</td>
<td>80.1</td>
<td>82.3</td>
</tr>
<tr>
<td>4</td>
<td>83.6</td>
<td>82.7</td>
<td>80.0</td>
<td>88.3</td>
<td>84.7</td>
<td>85.1</td>
</tr>
<tr>
<td>5</td>
<td><strong>94.6</strong></td>
<td>87.5</td>
<td>93.8</td>
<td>95.7</td>
<td><strong>91.8</strong></td>
<td><strong>97.2</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Order</th>
<th>SD-CD-D</th>
<th></th>
<th></th>
<th>SD-CD-D</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NAR-NN</td>
<td>NARX-NN</td>
<td></td>
<td>NAR-NN</td>
<td>NARX-NN</td>
<td></td>
</tr>
<tr>
<td></td>
<td>EPI</td>
<td>FCI</td>
<td>PSI</td>
<td>EPI</td>
<td>FCI</td>
<td>PSI</td>
</tr>
<tr>
<td>1</td>
<td>52.2</td>
<td>65.8</td>
<td>65.6</td>
<td>55.9</td>
<td>66.8</td>
<td>69.6</td>
</tr>
<tr>
<td>2</td>
<td>57.6</td>
<td>69.3</td>
<td>70.3</td>
<td>70.9</td>
<td>74.2</td>
<td>74.1</td>
</tr>
<tr>
<td>3</td>
<td>68.6</td>
<td>71.7</td>
<td>72.8</td>
<td>75.5</td>
<td>78.6</td>
<td>78.7</td>
</tr>
<tr>
<td>4</td>
<td>78.5</td>
<td>80.4</td>
<td>80.0</td>
<td>89.4</td>
<td>81.6</td>
<td>84.6</td>
</tr>
<tr>
<td>5</td>
<td>82.5</td>
<td><strong>86.2</strong></td>
<td>81.5</td>
<td>90.9</td>
<td>91.8</td>
<td><strong>97.2</strong></td>
</tr>
</tbody>
</table>

A comparison of average accuracy between speaker dependent and speaker independent cases is given in Figure 6-20. The NAR-NN were on average 5.5% and 3.5% more accurate for SD-CD-D and SD-CD-ND compared to the corresponding speaker independent models (SI-CD-D/SI-CD-ND). The NARX-NN models were 5.3% and 3.0% better with SD-CD-D and SD-CD-ND compared to SI-CD.
This indicates that both speaker and class dependence improves the overall prediction due to variations in characteristics of different speakers. To achieve the best results an individual should be trained on their own conversational pattern. A generic model of speaker independence reduces the performance of affect prediction.

6.4.2.6 Results comparing ANFIS with external output and without (EXP6)

The NAR-ANFIS and NARX-ANFIS models, introduced in Section 6.2.2.2, with grid partitioning used for FIS initialization as the optimal technique [41][44][110]. Only speaker dependent class dependent modeling (SD-CD) was investigated, based on the results from the NN experiments (i.e. EXP1-EXP5).

<table>
<thead>
<tr>
<th>Order</th>
<th>SD-CD-D</th>
<th></th>
<th>SD-CD-ND</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NAR-ANFIS</td>
<td>NARX-ANFIS</td>
<td>NAR-ANFIS</td>
<td>NARX-ANFIS</td>
</tr>
<tr>
<td></td>
<td>EPI FCI PSI</td>
<td>EPI FCI PSI</td>
<td>EPI FCI PSI</td>
<td>EPI FCI PSI</td>
</tr>
<tr>
<td>1</td>
<td>62.8 63.3 61.9</td>
<td>63.7 67.5 65.1</td>
<td>1</td>
<td>52.6 54.8 56.4</td>
</tr>
<tr>
<td>2</td>
<td>68.1 68.9 67.0</td>
<td>72.6 76.4 74.2</td>
<td>2</td>
<td>63.2 63.5 64.9</td>
</tr>
<tr>
<td>3</td>
<td>73.4 74.6 72.2</td>
<td>83.0 85.3 82.1</td>
<td>3</td>
<td>73.8 72.3 73.5</td>
</tr>
<tr>
<td>4</td>
<td>88.9 87.9 82.7</td>
<td>90.5 91.6 88.5</td>
<td>4</td>
<td>77.3 79.7 76.7</td>
</tr>
<tr>
<td>5</td>
<td><strong>95.9</strong> 93.3 89.7</td>
<td><strong>97.7</strong> 95.9 94.6</td>
<td>5</td>
<td>83.9 <strong>88.8</strong> 86.0</td>
</tr>
</tbody>
</table>

The average NAR/NARX ANFIS accuracies are given in Table 6-9 for depressed (left) and non-depressed (right) setups. The 5th order was the most accurate order for all cases. NAR-ANFIS attained 95.9% and 88.8% for SD-CD-D and DS-CD-ND and similarly NARX-ANFIS has 97.7% and 97.4% for SD-CD-D and DS-CD-ND.

The average improvement using NARX, across topics and depressed status, are 5.9%, 6.9%, 8.6%, 7.1% and 5.9% for order 1-5 compared to NAR. Generally SD-CD-D was more accurate suggesting depressed children have more predictable affect.
6.4.2.7 Comparison of NAR prediction implemented with ANFIS and NN (EXP7)

To summarize, the NAR/NARX models were implemented with two techniques: NN and ANFIS. The initial NN experiments showed speaker and class dependence (SD-CD) was optimal and used for ANFIS. The overall average accuracy, across topic and order, is given in Figure 6-21 for NAR-NN, NARX-NN, NAR-ANFIS and NARX-ANFIS for SD-CD-D and SD-CD-N. The results showed ANFIS was generally better than NN, except for the SD-CD-ND NAR in which NN outperformed ANFIS.

![Figure 6-21 Average accuracy (%), across order and interactions, for the Speaker Dependent Class Dependent (SD-CD-D and SD-CD-N) NAR and NARX implemented with NN and ANFIS](image)

6.5 Summary and Conclusions

This chapter presented RW and NAR/NARX based approaches for affect prediction of four constructs generated from LIFE annotations (ORI-DB). NAR/NARX learned patterns from dataset and was computationally expensive. The IRW used IC to weight the speakers where as the NAR/NARX used NN/ANFIS parameters as the weighting.

The RW experiments showed higher orders improve affect forecasting abilities and further improved using the newly proposed IRW utilizing IC from the EIM. The presented 2D RW is capable of predicting affect and provides a visual representation of emotional flow with observed differences in depressed and non-depressed subjects.
Preliminary experiments of NN implemented NAR/NARX compared speaker and class dependence and independence and consistently showed dependencies helped prediction. NARX implemented with NN/ANFIS improved the comparable NAR indicating the importance of external factors in prediction. The fuzzy interference system in ANFIS improved model capabilities compared to the NN.

The forecasting methods showed affect constructs in depressed was generally easier to predict than control subjects. This could be due to more predictable nature in depression including long periods of sadness and reduced emotional expression [449].

In both the random walk and non-linear auto-regression approaches the affect prediction accuracy improved considerably as the historical context of past states was increased. The explanation of this could be due to recurring patterns of affect or that the affect states change rapidly between observations. The latter suggestion would be most probable considering the nature of emotions typically change within seconds compared to mood that can last minute, hours or days and lasting longer is an affect disorder[507][508][509]. This concept is evident in the distribution of the annotated sequences discussed in Section 5.2.1 that showed the average duration of consecutive sequences, for all conversations, topics and gender lasted between 3-6 seconds depending on the construct.

The 2D random walk showed the possibility of tracking a conversation based on emotional transition patterns and emotional drifts to analyze a subject’s depressive state. This could eventually be used to model ideal interactions (i.e. non-depressed) and learn to eliminate undesirable interactions (i.e. depressed) and be used to test the effectiveness of treatments and behavioral therapies.
Chapter Seven:

CLASSIFICATION OF DEPRESSION USING ACOUSTIC FEATURES AND CLASSIFICATION SYSTEM

7.1 Preview

The methodology of this chapter is outlined by the flow diagram in Figure 7-1, which describes the proposed framework for detecting adolescent depression from conversations with parents. The ORI-DB used for this task was explained in Chapter Three: using a subset of dyadic conversation with a corpus from Section 5.2. The motivation of repeating existing acoustic depression detection approach was to have a direct comparison to the new adolescent depression detection CMS approach introduced in Chapter Five with the corresponding annotations.

The first main stage was speech pre-processing resulting in cleaned voiced speech segments, explained in Section 7.2. Next, acoustic features were extracted, from both parents and adolescents, described in Section 7.3. Then SVMs were used, explained in Section 7.4, to train and test models of depressed and control classes.

The experimental setups and procedures are discussed in Section 7.5 and the evaluation methods in Section 7.6. The experimental results comparing features and combinations are given in Section 7.7 with discussions and conclusions in Section 7.8.
7.2 Pre-processing

The pre-processing stage involved down sampling the speech signal and analysis to remove background noise and cross talk. The cleaned signal was normalized, framed into overlapping windows and a VAD used to extract voiced and remove silent and unvoiced segments. Details of these tasks are discussed in the following sub-sections.

7.2.1 Decimate Sampling rate

ORI-DB was recorded with a 44kHz sampling frequency and has redundant frequency components. The majority of speech energy is below 8kHz and is a useable frequency range in speech applications [282]. Memory requirements and processing cost are a concern with a large amount of data and a solution is down sampling.

An 11kHz sampling frequency was chosen to match past acoustic depression detection studies using ORI-DB [107][108][138][140][147]. This ensures all frequency components are covered with maximum resolvable frequency of 5.5kHz according to Nyquist theory. Down sampling causes aliasing distortion, which misinterprets higher frequency components, and solved with a 5.5kHz low-pass filter.

The sampling frequency was decimated (reduced) by a integer factor, $M$, as follows:

1) Remove high-frequency signal components using digital LPF

2) Down-sample the filtered signal by factor $M$
7.2.2 Background noise removal

The audio rerecording contains general background noise including ambient noise, which inhibit the audible sources. A common technique for removing the noise and improving speech quality is spectral subtraction (SS) introduced by Boll [123]. SS involves estimating the noise in the signal using an interval with no speech, normally at the start of the recording, and assumes the noise does not change significantly.

![Diagram of spectral subtraction process](image)

Figure 7-2 Spectral Subtraction process used to estimate the clean speech signal, \( x'(m) \), from the noisy speech, \( y(m) \), using an estimation of the noise, \( n(m) \).

A noisy signal, \( y(m) \), can be considered as the sum of the noise, \( n(m) \), and the desired signal, \( x(m) \). In the frequency domain the noisy signal, \( Y(f) \), is signified by the Fourier transforms of the desired signal, \( X(f) \), and estimated noise, \( N(f) \), as follows:

\[
Y(f) = X(f) + N(f)
\]  

(7.1)

The FFT is performed on the noisy signal frames windowed with a Hamming function and to alleviate effects of discontinuities and spectral leakage. The estimated noise is subtracted from the noisy signal, in the frequency domain, to obtain the estimate of the signal source, \( X'(f) \) as illustrated by Figure 7-2.

\[
|X'(f)|^b = |Y(f)|^b - \alpha |N(f)|^b
\]  

(7.2)

SS is defined by equation (7.2), where \( \alpha \) is the amount of noise subtracted, \( b \) denotes either magnitude \((b=1)\) or power \((b=2)\) spectral subtraction and the noise is estimated from a time-averaged noise spectrum, from \( N \) frames, as following:
\[
|N(f)|^b = \frac{1}{N} \sum_{n=1}^{N} |N_n(f)|^b
\]

(7.3)

For a detailed description on the methodology of spectral subtraction there is a vast amount of literature including the study by Boll [123]. In this application the spectral subtraction was implemented in Adobe Audition.

7.2.3 Source Separation

A common issue in signal analysis is interference of external sources on the desired signal required. During data collection of the ORI-DB each speaker was supplied with their own microphone, but due proximity of the speakers there was crossover between microphones, the speech of both participants were mixed into each stereo channel.

The ORI-DB was recorded under experimental condition and contains an insignificant amount of cross talk compared to the total length of recording. The sections with cross talk could be discarded although these segments are still important especially in parent-adolescent conversations [82]. Therefore, the simultaneous speech segments were processed to remove the cross talk and separate each channel.

![Diagram of cross-talk](image)

Figure 7-3 Diagram of cross-talk from two speakers (sources), \(s_n(t)\), that are involved in the mixing process defined by a weight, \(h_{nm}(t)\), creating mixed signals \(s_{nm}(t)\), that are summed at the microphone, \(x_n(t)\).
The problem of cross talk is illustrated by Figure 7-3 showing a system of two microphones, \(x_1(t)\) and \(x_2(t)\), and two speakers, \(s_1(t)\) and \(s_2(t)\). The microphone signals are recorded as mixtures of both the signal sources (speakers). The microphone signal (mixture) can be expressed as a weighted combination of the speakers (sources), such that the system of linear equations is given by equation (7.4), where \(h_{11}, h_{21}, h_{21}\), and \(h_{22}\) are parameters denoting the weighting of the sources.

\[
\begin{align*}
x_1(t) &= h_{11}s_1(t) + h_{12}s_2(t) \\
x_2(t) &= h_{21}s_1(t) + h_{22}s_2(t)
\end{align*}
\] (7.4)

In general, \(M\) observed linear mixtures (microphone signals), \(x_1, x_2, ..., x_M\) of \(N\) independent components (speaker sources), \(s_1, s_2, ..., s_N\), are expressed as the weighted sum of each source given by equation (7.5). Where the weights are defined as \(h_{mn}\) such that \(m\) is mixture index and \(n\) is source index. The aim is to recover the original speech sources, \(s_n(t)\), using the observed recordings, \(x_m(t)\).

\[
x_m(t) = \sum_{n=1}^{N} h_{mn}s_n(t), \quad m = 1, ..., M
\] (7.5)

The difficulty is that the weights are unknown, which is a blind source separation (BSS) problem, such that the ‘source’ needs to be separated being ‘blind’ to the mixing matrix (weights). BSS can be solved with various approaches including independent component analysis (ICA), principal component analysis (PCA) and factor analysis [494]. ICA is capable of finding sources when others fail, is more powerful compared to PCA and is a widely used approach [490].
ICA is a generative model that assumes observed multivariate data is a linear mixture of unknown latent variables and an unknown mixing matrix. ICA estimates sources the latent variables that are independent components (sources) of observed data and assumed to be statistically independent and non-Gaussian.

It is convenient to use a matrix notation of a statistical “latent variable” model by rewriting equation (7.5) as equation (7.6) [493]. Such that s are latent variables as a vector with elements being the sources \( s_n \), the weights \( h_{mn} \) are the elements of the mixing matrix, \( H \), and \( x \) is the observed data with mixtures \( x_m \) as the elements.

\[
x = HS
\]  

(7.6)

ICA estimates the mixing matrix, \( H \), by assuming statistical properties of the source, such as statistical independence and non-Gaussian, which are realistic assumptions [489].

The mixing matrix is inversed, \( W \), which is used to find the sources (independent components) from the mixtures of the components as follows:

\[
s = WX
\]  

(7.7)

In this application the cross-talk segments are processed using the FastICA algorithm implemented with fast fixed-point algorithm ICA using the MATLAB package [491][492][489]. FastICA is a popular ICA algorithm due to fast operation, computational efficiency in large-scale applications and statistical robustness [491].

### 7.2.4 Voiced Speech Windows

The cleaned signal was normalized and windowed with a Hamming function into 25ms with 50% overlap. The length was chosen to cover an entire periodic cycle of speech and capture the fundamental frequency, normally occurring within 2-12ms [499]. The windows were analyzed to extract voiced speech and remove unwanted unvoiced and silent segments using VAD from Section 4.3.1 and described in [184].
7.3 Acoustic Feature Extraction

As discussed in Section 2.2.2-2.2.4 and 2.3 speech production, occurring through the glottis, vocal folds, vocal tract and mouth, is physiologically equated to acoustic features related to emotion and depression. Depression studies generally follow procedures that involve grouping acoustic features into categories that relate to the human speech production model [140][139][147][106][105]. The categories chosen relate to physiological and perceptual components that characterize speech as follows:

- Prosodic (P)
- Spectral (S)
- Cepstral (C)
- Teager based (T)
- Glottal (G)
- Glottal Waveform (GW)

These categories provide characteristics related to a range of speech production concepts. Glottal related features are linear physiological components related to the glottal excitation source. Prosodic and spectral features are physiological components related to linear speech production model through the glottis, vocal folds, vocal tract and filtering. TEO features are physiologically based parameters derived from nonlinear model of the vocal tract. Cepstral features are related to the linear speech production with an adaptation to a perceptual aspect.
A summary of the acoustic speech feature categories are outlined in Table 7-1, defining the sub-category features and the number of coefficients. The investigation of features provided insight into contributions of acoustic parameters in relation to depression. Detailed methodologies and explanations for each subcategory feature are given in the following subsections.
7.3.1 Prosodic

In speech acoustics, prosodic features are connected with subjective measures of speech quality related to loudness, pitch perceptions and speaking rate. Prosodic characteristic are perceivably changed in depressed speakers [104], often being described as monotonous, flat, dull or lifeless [293].

Prosodic features are widely used in speech, emotion, stress and more recently success in depression [438][140][109][106][139][108]. The features used in this research include fundamental frequency ($F_0$), log energy ($\text{LogE}$), shimmer and jitter.

7.3.1.1 Fundamental Frequency ($F_0$)

The fundamental frequency, $F_0$, of vocal fold vibration can be estimated with time or frequency domain approaches: autocorrelation, cepstrum and average magnitude difference [259]. Previous studies have determined autocorrelation was optimal in acoustic analysis [140][259]. In this application $F_0$ was estimated using the autocorrelation method, described in [214], which estimates the similarity between a frame, $x(n)$, and its delayed version, $x(n+\tau)$. For a periodic signal there is peak correlation at 0 lag and at delays, $\tau$, corresponding to multiples of the pitch period.

\[
ACF(\tau) = \sum_{n=0}^{N-1-\tau} x(n)x(n + \tau) \tag{7.8}
\]

The autocorrelation function (ACF) for each frame is defined by equation (7.8), where $N$ is the frame size, $x(n)$ and $\tau$ is the lag. The time lag, $\tau > 0$, at the first ACF peak corresponds to the pitch period ($1/ F_0$) and converted to $F_0$. The search space was restricted to 40-1000 Hz as recommended by previous studies [147][140].
7.3.1.2 Log energy (LogE)

The speech waveform energy was generated on a frame-by-frame basis using the log-energy parameter, which is correlated to perceived loudness [334]. The LogE for a given frame was the logarithm of the sum of squared amplitudes, given by equation (7.9) where \( n \) is the sample index and \( N \) is the number of samples.

\[
LogE = \log \sum_{n=1}^{N} x^2(n) \tag{7.9}
\]

7.3.1.3 Jitter

Jitter is the variation in the fundamental frequency, \( F0 \), from cycle-to-cycle (short-term) fluctuations. Jitter was based on the average relative difference between the \( F0 \) of consecutive frames as described by equation (7.10), where \( F0_i \) is the fundamental frequency of frame \( i \) and \( M \) is the total number of frames.

\[
Jitter = \frac{1}{M-1} \sum_{i=1}^{M-1} |F0_i - F0_{i+1}| \frac{1}{M} \sum_{i=1}^{M} F0_i \tag{7.10}
\]

7.3.1.4 Shimmer

Shimmer is a measurement of the cycle-to-cycle variation in the speech waveform peak-to-peak amplitude. Shimmer was defined as the average relative difference between the peak-to-peak amplitude, \( A \), of consecutive frames given by equation (7.11), where, \( i \) is the index and \( M \) is the total number of frames.

\[
Shimmer = \frac{1}{M-1} \sum_{i=1}^{M-1} \left| 20 \log \left( \frac{A_{i+1}}{A_i} \right) \right| \frac{1}{M} \sum_{i=1}^{M} A_i \tag{7.11}
\]

Jitter and shimmer both measure cycle-to-cycle variation in speech and have been successfully applied towards depression detection [140][108][422]. Both variables have been generated using the process described by Childers [184].
7.3.2 Spectral

Spectral features are related to frequency components of speech characterizing speech spectrum properties. Spectral features have been previously used in acoustic depression detection and mostly outperform prosodic features [139][108][140][424].

In this research spectral features calculated include: flux, centroid, entropy, roll-off, formants (bandwidth) and power spectral density (PSD) with sub-bands and ratios.

7.3.2.1 Formants

Formants are spectral peaks of the spectrum and a measure of acoustic resonance frequencies of the vocal tract[124][338]. Formants deliver important information relating to acoustic speech characteristics essential in analysis [125][450]. Formants are significantly distinguishable in speech of depressed subjects [432][390] and compared to after treatment [156] and have been used in depression classification.

Formants can be estimated as the peaks in the spectral envelope of the vocal tract signal, modeled as an all-pole filter using a $13^{th}$ order linear prediction (LP) filter. The first five formants were obtained from the roots of the polynomial LP predictor (positive imaginary parts). The poles, $p_i$, of the transfer function correspond to formants of the vocal tract spectrum. The $i^{th}$ formant, $F_i$, and corresponding bandwidth, $BW_i$, are calculated in the $z$-plane as follows:

\[
F_i = \frac{F_s}{2\pi} \left( \tan^{-1} \frac{\text{Im}(p_i)}{\text{Re}(p_i)} \right) \quad (7.12)
\]

\[
BW_i = -\left( \frac{F_s}{\pi} \right) \ln(|p_i|) \quad (7.13)
\]
7.3.2.2 Power spectral density

The single-sided power spectral density (PSD) has been used to discriminate between speech of control and depressed adults [109]. In this application the PSD was estimated using a non-parametric approach via the Welch spectral estimator [306]. A 4096-point FFT generated the frequency spectrum with 50% overlapped Hamming windows and the PSD for the entire bandwidth given by equation (7.14).

\[
PSD_{dB} = 10 \log_{10}(PSD) \tag{7.14}
\]

The total power, \( P_{total} \), for 0-2000Hz frequency range was determined by calculating the area under the \( PSD_{dB} \) using trapezoidal numerical integration. The power calculation was repeated for spectral sub-bands, 0-500Hz, 500-1000Hz, 1000-1500Hz, and 1500-2000Hz, generating \( P_{sub-band} \). The ratios of powers of each sub-band to the total power of the whole bandwidth are given by equation (7.15)

\[
\text{ratio} = \frac{P_{sub-band}}{P_{total}} \tag{7.15}
\]

7.3.2.3 Spectral flux

Spectral flux measures the amount of change in the power spectrum \( (PS) \) of a signal between consecutive frames, \( i \) and \( i-1 \). The Euclidean distance measures the distance between the normalized power spectrums of consecutive frames as follows:

\[
Fx_{S} = \| |PS_{i}| - |PS_{i-1}| \| \tag{7.16}
\]

7.3.2.4 Spectral centroid

The center of a signal’s spectral power distribution is defined as the spectral centroid, \( C_{S} \), given by equation (7.17), as the weighted mean of the frequencies present in the signal. Where the power spectral magnitudes, \( PS(k) \), are used weight the corresponding frequency, \( f(k) \), with a total of \( K \) frequency bins.

\[
C_{S} = \frac{\sum_{k=1}^{K} PS(k)f(k)}{\sum_{k=1}^{K} PS(k)} \tag{7.17}
\]
7.3.2.5 Spectral entropy

Spectral entropy is a measure of the amount of information of a signal (Shannon’s information theory). Entropy is used to measure the spikiness of a spectral distribution and captures the formants and peaks in the signal. The main steps involved in calculating the spectral entropy for each frame was as follows:

- Calculate the power spectrum of signal \(X(w)\) using digital Fourier transform

- Calculate the power spectral density (PSD), equation (7.16), where \(N\) is the number of points in the spectrum and \(i\) defines the \(n^{th}\) frequency component

\[
PSD_n = \frac{1}{N} |S_n|^2
\]  
(7.18)

- The frequency components are normalized so the spectrum can be viewed as a Probability Density Function. The normalized power spectrum, \(NPS\), is given by equation (7.19) where \(n\) is the frequency component index and \(N\) the FFT length.

\[
NPS_n = \frac{PSD_n}{\sum_{n=1}^{N/2} PSD_n}
\]  
(7.19)

- Compute the entropy using equation (7.20) where \(NPS_n\) is the normalized power spectral density at the corresponding frequency bin.

\[
H = - \sum_{n=1}^{N/2} NPS_n \log_2 (NPS_n)
\]  
(7.20)
7.3.2.6 Spectral roll-off

In audio signals energy has a tendency to be lower at high frequencies, known as the spectral slope [121], characterized by the energy and frequency relationship using spectral roll-off. The spectral roll-off is defined as frequency, $f_R$ that is below a set percent, $k$, of the power spectrum resides. In theory and past studies the percentage, $k$, has mostly been set as 75%, 80%, 85% or 95% [122][147][140][108].

\[
\text{argmin}_{f_R \in \{1, \ldots, N\}} \sum_{n=1}^{f_R} PS(n) \geq k \sum_{n=1}^{N} PS(n) \quad (7.21)
\]

The spectral roll-off was calculated according to equation (7.21) where $N$ is the number of frequency bins, $n$ is the frequency bin index, $PS(n)$ is the corresponding spectral magnitude and $f_R$ is the spectral roll-off. The spectral roll-off defines the frequency below which a percentage, $k$, of the total spectral energy was accumulated.

Some studies have considered a range of spectral roll-offs, generally within a limited range [399][404][424]. This research replicated this idea with higher resolution using 5% intervals of $k$ values from 5% to 95%.
7.3.3 Cepstral (MFCC)

Cepstral features, especially based on the power cepstrum, are known to be effective in human speech analysis [350][351]. The advantage of MFCC is due to the cepstrum being characterized by the perceived human auditory system [56][57][58]. MFCC are widely recognized as a suitable acoustic representation for speech [352][354], speaker [353][59] and emotions recognition tasks [356][342][366][450][19].

More recently MFCC have been successfully applied towards psychology applications showing potential in classification of suicide risk of near-suicidal patients[484][485]. Newer studies have shown encouraging results in depression detection [108][138][140][427][421][422] and in certain cases has outperformed other features such as spectral, prosodic and TEOs [107][424][434][435]

Adolescent depression detection is based on MFCC extracted from adolescent speech and furthermore, for the first time, their parent’s speech. Details on the derivation of the MFCC is identical to that of the AER experiments in Section 4.4.1

7.3.4 TEO critical-band autocorrelation envelope (TEO-CB-Auto-Env)

TEO features have shown good performances in stress[20][238][55][19] and emotion recognition [19][64][189][235]. Emotional states such as anger and stress, causes vortices that provide additional excitation signals [68]. Studies have suggested the airflow in speech production show distinguishable physical manifestations in depressed subjects and have significant effect on the vocal folds [462][463].

Therefore the use of TEO has been extended towards affect disorders, mental disorders and pathology [461]. The most recent studies in depression detection have improved results, compared to prosodic and spectral features, utilizing TEO [107][108][140][147]. In this application adolescent depression in detected using TEO_CB_Auto_Env, described in Section 4.4.2, suggested in [55].
7.3.5 Glottal Features

The glottal waveform represents air-flow generated during vocal fold opening and closing. Glottal pulse shape and related features are reported to play an important role in speech production and have links to psycho-physiological mechanisms, which changes speech phonation. This has steered studies to explore glottal features in speech [195], speaker [389][388][208], stress [401][194] and emotion [63][64][197][168][391] recognition improving upon spectral and prosodic features.

Glottal features have shown strong correlation, high importance and provided excellent results in depression detection, mostly outperforming other features [92][106][65]. Studies have indicated the addition of glottal features to other features (spectral and prosodic) enhances depression detection [105][140][106].

![Figure 7-4 Diagrams](image)

Figure 7-4 Diagrams of (a) top- Time-domain speech frame (b) middle- Glottal waveform estimate using TKK Aparat toolbox (c) bottom- Glottal flow derivative. Including the relevant phases and parameters required to generate the glottal time and frequency parameters.
The glottal waveform was estimated using the TKK Aparat glottal inverse filtering toolbox [185]. The inverse filtering uses an IAIF algorithm [66], based on DAP, previously described in Section 4.3.2. Quantitative analysis of the glottal waveform (pulse) has been examined in the glottal time (GTD) and frequency (GFD) domain, generated with TKK Aparat toolbox [185]. Before providing detail on parameters (GTD/GFD) background knowledge is delivered on the glottal waveform.

Important timing instances include the duration of the: primary and secondary opening phase [466], $T_{o1}$ and $T_{o2}$, the closing phase, $T_c$, and the period of the glottal cycle, $T$. These values are essential to calculate timing and frequency parameters, GTD and GFD, described in the following subsections and more detail in [185][140].

### 7.3.5.1 Glottal Time Domain (GTD)

The glottal timing (GTD) parameters represent characteristics related to the amplitude, timing and duration of the vocal folds opening and closing phases. In this application nine GTD parameters are extracted including:

- **Open Quotients:** Ratio of the primary ($T_{o1}$) /secondary ($T_{o2}$) opening phase to the glottal cycle duration, $T$. Denotes amount of time the folds are open in glottal cycle

\[
OQ_1 = \frac{T_{o1} + T_c}{T} \quad (7.22)
\]

\[
OQ_2 = \frac{T_{o2} + T_c}{T} \quad (7.23)
\]

- **Approximation of the Open Quotient:** Approximate opening to OQ with an ideal LF pulse. Where $U_{ac}$ is the glottal waveform peak-to-peak amplitude and $\Delta U_{\text{max}}$ and $\Delta U_{\text{min}}$ amplitudes from the derivative of the glottal waveform.

\[
OQ_a = U_{ac} \left( \frac{\pi}{2\Delta U_{\text{max}}} + \frac{1}{\Delta U_{\text{min}}} \right) f_0 \quad (7.24)
\]
• **Quasi-Open Quotient**: Time the open phase duration that is 50% above the peak-to-peak amplitude of the glottal flow

\[ QOQ = \frac{T_{q(50)}}{T} \]  

(7.25)

• **Speed Quotients (Skewness)**: Ratio of duration of primary \( T_{o1} \)/secondary \( T_{o2} \) opening phase to the closing phase, \( T_c \). Represents the portion of time in each cycle with the folds moving outward divided by time folds moving inward. It describes how skewed the pulse is (how far from symmetric)

\[ SQ_1 = \frac{T_{o1}}{T_c} \]  

(7.26)

\[ SQ_2 = \frac{T_{o2}}{T_c} \]  

(7.27)

• **Closing Quotient**: Ratio of peak-to-peak amplitude, \( U_{ac} \), of glottal flow to minimum peak of pulse derivative, \( \Delta U_{min} \).

\[ CQ = \frac{U_{ac}}{\Delta U_{min}} \]  

(7.28)

• **Amplitude Quotient**: Ratio of timing duration of the closing phase, \( T_c \), to the glottal cycle length, \( T \).

\[ AQ = \frac{T_c}{T} \]  

(7.29)

• **Normalized Amplitude Quotient**: Normalized \( AQ \) by dividing it by the glottal cycle duration, \( T \).

\[ NAQ = \frac{AQ}{T} \]  

(7.30)

The quotient parameters define how the glottal waveform is shaped and determine the loudness and timbre of the speech signal. A smooth pulse with gradual airflow changes, large open quotient and skewing quotient, tends to produce 'fluty' sound. A less-smooth waveform with sudden airflow changes, large skewing quotient or small open quotients, results in more high frequencies and a 'brassy' timbre.
7.3.5.2 Glottal Frequency Domain (GFD)

The glottal frequency (GFD) parameters were extracted from the glottal waveform spectrum including the following three parameters generated from the harmonics:

- **Difference between Harmonics**: Difference between the 1st and 2nd harmonics, \((H_1, H_2)\) measured in decibels of the glottal flow power spectrum

\[
DH = H_1 - H_2
\]  
(7.31)

- **Harmonic Richness Factor**: Ratio of the sum of harmonic magnitudes, \(H_k\), to the magnitude of the first harmonic, \(H_1\)

\[
HRF = \frac{\sum_{k>2} H_k}{H_1}
\]  
(7.32)

- **Parabolic Spectral Parameter**: Fits a 2nd order polynomial to the glottal flow spectrum on a logarithmic scale with the PSP extraction outlined in [467].

7.3.5.3 Glottal MFCC and Glottal TEO (G-MFCC and G-TEO)

MFCC and TEO features were extracted from the estimated glottal waveform, detailed in Section 4.3.2, following the same method as outlined in Sections 7.3.3 and 7.3.4, generating glottal MFCC and glottal TEO (G-MFCC and G-TEO).

These features were chosen based on AER studies that showed improvement using glottal waveform TEO [238] and MFCC [342][367]. Minimal research on features extracted from the glottal waveform has been applied towards depression and to our knowledge have been limited to G-TEO [65] and G-MFCC has not been used.
7.4 Modeling and Classification Setup

There are many machine learning approaches including commonly used classifiers, GMM [364][207], SVM [61][407][409][207] or NN that have been utilized in many pattern recognition problems including depression detection applications.

A disadvantage of GMM is the unclear choice of the number of mixtures and that it fails in higher dimensional problems due to computational demands. In methods such as Naïve Bayes all data points influence the optimal model in contrast SVMs only use points close to the decision boundary (support vectors) [411].

SVMs are beneficial for several reasons: learns simple linear and complex non-linear models (kernel), robust to outliers and noise (slack variables), robust to high-dimensional data and limited samples, does not suffer from curse of dimensionality, avoids over fitting and solves global minimum efficiently.

SVMs are often considered the state-of-the-art classifier and have advantageous properties with generalization capabilities for binary problems [406][408]. Generally, SVM are able to classify binary problems more accurately than other classifiers [309]. The SVM was rationalized as the ideal choice due to established theoretical foundation and past success [287][288][289][408][61][407].

Depression detection used SVM classification implemented with the LIBSVM toolbox [291] with a radial basis function (RBF) kernel and sequential minimization optimization (SMO) [292]. A 3-stage grid-search, on 10% of the dataset, was used to optimize the hyper-parameters ($C$, $\gamma$) based on increasing the CV accuracy. Each grid-search stage reduced the search space around the current optimal parameters and fine-tuned with increasing precision: big, medium and small scales. The remaining 90% of data was segmented into 80% for training to learn the $w$ and $b$ parameters and the remaining 20% for testing with 3-fold CV and used the optimal hyper-parameters.
7.5 Experimental setup for acoustic classification of depression

The experiments were based on the methodology explained in Sections 7.2, 7.3 and 7.4, with an investigation on five common acoustic feature categories in depression detection studies: TEO-based (T), cepstral (C), prosodic (P), spectral (S) and glottal (G), and the introduction of a sixth category related to the glottal waveform (GW).

The S category was investigated further by optimizing a range of spectral roll off parameters, through mRMR, to improve depression detection. The features were examined as individual feature set sub-categories, collective categories and combinations to distinguish acoustic feature that improve depression classification.

The investigation in this section used the speech signals from the ORI-DB corpus, described in Section 3.3.5, as a direct comparison to the CMS depression detection approach from Chapter Five using the corresponding annotations.

The dataset was divided so at least 20% of the speakers were used for testing, and the rest was used for training the models. Due to the size difference of each class’s data a random selection was removed from the larger class to ensure a balanced dataset with 50% for a random chance (binary case). The experiments are briefly outlined below with more details in the results and discussion sections.

- **EXP1**: Determined the optimal correlates of adolescent depression using statistical analysis (MANOVA and ANOVA) on the acoustic feature categories and sub-categories for the parents and adolescents.

- **EXP2**: Examined the role of gender in depression detection by comparing gender independent (GIM) and gender dependent modeling (GDM) in adolescent depression classification using $F0$ as a basic feature [124][496][500]. This was conducted due to evidence that depression and speech differ between genders, especially during adolescence [213].
EXP3: Determined the most important range of spectral roll-off coefficients, from the spectral category, for depression discrimination based on mRMR. The ranking of the spectral roll-offs was used in a filter selection filter approach to iteratively optimize the SVM depression classification.

EXP4: Compared performance of multiple individual acoustic features sets from prosodic and spectral (including the roll-off from EXP3) sub-categories using an SVM with GDM and GIM found from EXP2.

EXP5: Similar to procedures in past studies[140][105], the acoustic features were grouped into two categories of prosodic (P) and spectral (S) using only significant features determined from EXP1. This is compared a new spectral category with the addition of the proposed optimized roll-offs found in EXP3.

EXP6: Investigated the performance of relatively new features towards depression classification, [140][105][107][108], using higher dimension feature sets comparing MFCC and TEO-CB-Auto-Env.

EXP7: Examined the noise robustness of MFCC and TEO-CB-Auto-Env by determining the effect varying levels of additive white Gaussian noise (AWGN) has on depression classification performance compared to EXP6.
- **EXP8**: Assessed glottal category (G) performance and individual subcategories with glottal time (GTD) and glottal frequency domain (GFD)

- **EXP9**: Based on the well-known importance of the glottal waveform (GW) both MFCC and TEO, from EXP6, were extracted from the glottal waveform creating G-TEO used in [65] and G-MFCC new to depression application.

- **EXP10**: As suggested by previous studies the addition of the glottal features enhances classification compared to stand alone features[105][140][106]. This is investigated by fusing the glottal category (G) from EXP8 with P and S categories from EXP5 and C and T from EXP6. Similarly, it is proposed to fuse a combination of G and the new glottal waveform category (GW) from EXP9.

- **EXP11**: Due to the highest performance of the Spectral category (S), compared to all six f categories (P, S, C, T, G, GW) based on EXP5, EXP6, EXP8 and EXP9, an experiment was conducted to determine if S combined with the remaining categories could improve on the stand-alone feature.

- **EXP12**: Repeated, EXP5, EXP6, EXP8 and EXP9 using the speech from the parents to determine using the parent’s categorical (S, P, C, T, G, GW) acoustic features can detect depression in their offspring.
7.6 Evaluation Methods

The acoustic depression detection systems were assessed based on the sensitivity, specificity and accuracy, given by equations (7.33), (7.34) and (7.35) [167], providing quantitative measure of depression detection precision.

\[
\text{Sensitivity} = \frac{TP}{TP + FN} \times 100\% \quad (7.33)
\]

\[
\text{Specificity} = \frac{TN}{TN + FP} \times 100\% \quad (7.34)
\]

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (7.35)
\]

Depression detection studies have shown a subject-based approach improve accuracy compared to utterance/frame level [454][140][65]. Subject-based evaluations are only necessary and helpful in datasets with relatively few subjects [92]. Therefore, it was chosen to calculate the true positive (TP), false positive (FP), false negative (FN) and true negative (TN) parameters on a frame-based case as follows:

- TP: number of samples correctly classified as depressed
- FP: number of non-depressed samples misclassified as depressed
- TN: number of samples defined correctly classified as non-depressed
- FN: number of depressed samples misclassified as non-depressed

The objective is to detect depression so it is more important to identify depressed subjects (higher sensitivity) rather than correctly determine non-depressed (higher specificity). To achieve this requirement a conditions must be met to ensure a high sensitivity to specificity ratio \((s_e/s_p>1)\) but without skewing the classes \((s_e/s_p < 2)\).
7.7 Experimental results

7.7.1 Statistical Results- MANOVA and ANOVA (EXP1)

The acoustic feature categories, described in Section 7.3, have been examined for statistically significant speech features. Multivariate analysis of variance (MANOVA) determined features with significant correlation in a pairwise comparison of depressed and control cases. MANOVA was conducted on separate feature categories, as it is suboptimal to combine all features in theory [460] and practice [140]. MANOVA were assessed using Wilk’s lambda statistical procedure and significant if \( p<0.05 \) and followed by ANOVA to analyze the significance of the sub-categories.

Table 7.2 MANOVA and ANOVA analysis comparing depressed and control subcategory features for male and female adolescents. Where “+” denotes significance (\( p<0.05 \)) and “-” denotes not significant.

<table>
<thead>
<tr>
<th>Category</th>
<th>Sub-Category</th>
<th>No. of Parameters</th>
<th>Male</th>
<th></th>
<th></th>
<th>Female</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>EPI</td>
<td>FCI</td>
<td>PSI</td>
<td>EPI</td>
<td>FCI</td>
<td>PSI</td>
</tr>
<tr>
<td>Prosodic (P)</td>
<td>F0</td>
<td>1</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>LogE</td>
<td>1</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>Shimmer</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Jitter</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Spectral (S)</td>
<td>Flux</td>
<td>1</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>Centroid</td>
<td>1</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>Entropy</td>
<td>1</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>Roll-Off</td>
<td>19</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>PSD</td>
<td>9</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>Formants and BA</td>
<td>10</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Cepstral (C)</td>
<td>MFCC</td>
<td>12</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>TEO (T)</td>
<td>TEO-CB-Auto-Env</td>
<td>17</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Glottal (G)</td>
<td>Glottal Time (GT)</td>
<td>9</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>Glottal Frequency (GF)</td>
<td>3</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Glottal Waveform (GW)</td>
<td>G-MFCC</td>
<td>12</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>G-TEO</td>
<td>17</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

Table 7-2 provides statistical significance of feature categories (MANOVA) and subcategories (ANOVA) for each gender and interaction. The statistical tests discriminate the depressed status of an adolescent with significance is denoted by “+” and non-significance, \( p>0.05 \), denoted by “-“. Every case showed statistical significance with the exception of shimmer and jitter from the Prosodic (P) category, consistent with past studies that found no significance in mean jitter [106][428].
7.7.2 Effectiveness of Gender Dependent vs. Gender Independent Models (EXP2)

Speech recognition tasks require gender dependence, due to acoustic differences present in male and female speech [235][234][172][473]. In addition psychological studies have concluded depression symptoms [420][471] and speech [100][213][82] differ in genders. Accordingly, studies have improved detection rates using gender dependent modeling [140][138][435][107].

In this preliminary experiment gender dependence was investigated using F0 as a starting point, chosen as a basic feature with clear and well recognized distinctions in male and female speech [124][496][500]. SVM were trained for gender independent model (GIM) using both male and females. In contrast the GDM were generated separately using only female (GDM-F) or male (GDM-M) subjects.

<table>
<thead>
<tr>
<th>Features</th>
<th>EPI</th>
<th>FCI</th>
<th>PSI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sens</td>
<td>Spec</td>
<td>Acc</td>
</tr>
<tr>
<td>GDM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>71.4</td>
<td>54.5</td>
<td>62.9</td>
</tr>
<tr>
<td>Female</td>
<td>72.5</td>
<td>59.7</td>
<td>65.9</td>
</tr>
<tr>
<td>GIM</td>
<td>61.9</td>
<td>49.2</td>
<td>55.6</td>
</tr>
</tbody>
</table>

Table 7-3 gives the SVM depression classification performance, for the GIM and GDMs (GDM-M and GDM-F). It was observed for all topics (EPI, FCI and PSI) that the GIM was consistently weaker than both GDMs. GIM was degraded by 7%, 8% and 8% compared to GDM-M and 10%, 14% and 25% compared to GDM-F for EPI, FCI and PSI respectively. Furthermore, it can be noted that it was easier to detect depression in females compared to males and PSI over the FCI and PSI.

The results suggest that it is necessary to continue on with gender dependent systems, which agrees with the current literature of gender and depression [420][107][471][213][82][100] and depression detection studies [140][138][435].
7.7.3 Optimize Feature Selection from Spectral Roll-off Coefficients (EXP3)

As previously stated in Section 7.3.2.6 the spectral roll-off is defined as the frequency, \( f_R \), which a specified proportion, \( k \), of the spectrum is within. In past studies the spectral-roll has been defined as the frequency that majority of the energy exists (i.e. 75%, 80%, 85%, 95%) [122][147][140][108] or a range of spectral roll-off values [399][404][424]. In this research the spectral roll-off has an increased resolution and range of parameters for \( k \) values of 5% to 95% with 5% increments.

Figure 7-5 Illustration showing the relative difference of the average spectral roll-offs between the non-depressed (ND) and depressed (D) subjects for a range of cut-off points (\( k=5\%-95\% \)) included for male and female adolescents in EPI, FCI and PSI

Examining the range of spectral roll-off is analyzed by comparing the spectral roll-off of each class as a relative difference of the average spectral roll-off between depressed and non-depressed for each ratio (\( k \)) given in Figure 7-5. Numerous findings are observed with clear distinctions in mean roll-off in relation to depression.

There was a large difference in the spectral roll-off frequencies, between D and ND, in reference to 30% and 80% of the energy distribution. Depending on the interaction and gender case it can be seen that around 5%-10% and 50% -80% intervals show minimal difference that match up with the \( p \)-values, given by Table 7-4, denoting statistical insignificance of the spectral roll-off between the D and ND.
The roll-offs below 55% are higher for ND and above 55% higher for D. The interoperation is that D compared to ND has less (more) than 55% of energy is concentrated below a lower (higher) frequency. This implies D subjects have a higher energy concentration in higher frequencies than ND and that ND has relatively more energy concentrated in the lower frequencies. In general energy has a tendency to be lower at high frequencies [121] and this is truer in the ND case.

Based on evidence of the importance of the entire spectral roll-off range it could be expected that this could provide additional information towards depression detection. Further investigation is necessary to optimize the feature set to remove the parameters that are irrelevant for depression classification. Instead of an exhaustive search of spectral roll-off combinations a less computationally expensive approach is filter based using a type of statistical feature ranking of a specified sub-set size.

### Table 7-4 ANOVA Analysis on Depressed and Control Spectral Roll-off Features for Male and Female Adolescents Where “<” Denotes p<0.001 and The Shaded Cells Indicate Significance (p<0.05)

<table>
<thead>
<tr>
<th>k(%)</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EPI</td>
<td>FCI</td>
</tr>
<tr>
<td>5</td>
<td>&lt;</td>
<td>0.812</td>
</tr>
<tr>
<td>10</td>
<td>&lt;</td>
<td>0.043</td>
</tr>
<tr>
<td>15</td>
<td>&lt;</td>
<td>&lt;</td>
</tr>
<tr>
<td>20</td>
<td>&lt;</td>
<td>&lt;</td>
</tr>
<tr>
<td>25</td>
<td>&lt;</td>
<td>&lt;</td>
</tr>
<tr>
<td>30</td>
<td>&lt;</td>
<td>&lt;</td>
</tr>
<tr>
<td>35</td>
<td>&lt;</td>
<td>&lt;</td>
</tr>
<tr>
<td>40</td>
<td>&lt;</td>
<td>&lt;</td>
</tr>
<tr>
<td>45</td>
<td>&lt;</td>
<td>&lt;</td>
</tr>
<tr>
<td>50</td>
<td>&lt;</td>
<td>&lt;</td>
</tr>
<tr>
<td>55</td>
<td>0.052</td>
<td>0.81</td>
</tr>
<tr>
<td>60</td>
<td>&lt;</td>
<td>&lt;</td>
</tr>
<tr>
<td>65</td>
<td>&lt;</td>
<td>&lt;</td>
</tr>
<tr>
<td>70</td>
<td>&lt;</td>
<td>&lt;</td>
</tr>
<tr>
<td>75</td>
<td>&lt;</td>
<td>&lt;</td>
</tr>
<tr>
<td>80</td>
<td>&lt;</td>
<td>&lt;</td>
</tr>
<tr>
<td>85</td>
<td>&lt;</td>
<td>&lt;</td>
</tr>
<tr>
<td>90</td>
<td>&lt;</td>
<td>&lt;</td>
</tr>
<tr>
<td>95</td>
<td>&lt;</td>
<td>&lt;</td>
</tr>
</tbody>
</table>
Feature optimization was achieved using Minimum redundancy and maximum relevancy (mRMR) approach, previously outlined in Section 5.9.6, as a filter based feature selection method [37]. mRMR ranks the feature parameters that best characterize the properties of the target (class) that discriminate between depressed and non-depressed class implemented with Mutual Information Quotient (MIQ) [38].

A second-stage wrapper follows the mRMR filter to optimize the feature subset by maximizing depression classification accuracy [38]. This was carried out by iteratively removing the lowest ranked coefficient and training 3-fold cross-validated SVM, summarized by Figure 7-6 and the following procedure:

1) Generate the mRMR feature selection and rank the spectral roll-off parameters based on the MIQ form most to least important

2) SVM classification of the selected feature coefficients with the lowest ranked removed one at a time

Figure 7-6 Framework of the feature selection method using a 2-stage mRMR filter to generate a ranking of top acoustic feature category used in a SVM wrapper stage to iteratively remove the lowest ranked feature to find the optimal accuracy

SVM depression classification accuracy for GDM-M and GDM-F are given by Figure 7-7 and Figure 7-8 using iteratively reduced subsets of spectral roll-off coefficients with mRMR feature selection strategy. The crosses signify the optimal feature subset with the highest accuracy. The best subset kept 14, 16, 17 (GDM-M) and 11, 10, 17 (GDM-F) of the top ranking coefficients for EPI, FCI and PSI.
FOR THE GDM-GDM Features SPECIFICITY USING A roll roll
Figure 7-7 Accuracy (%) of SVM GDM-M depression classification with 1 to 19 roll-off feature coefficients kept based on ranking of MIQ from the mRMR filter.

Figure 7-8 Accuracy (%) of SVM GDM-F depression classification with 1 to 19 roll-off feature coefficients kept based on ranking of MIQ from the mRMR filter.

Table 7-5 compares the highest depression classification accuracy to the entire roll-off set and best individual feature. For all cases the entire roll-off feature set inferior compared to the optimized feature subset for GDM-M by 9%, 5% and 2% and GDM-F by 3%, 7% and 5% in EPI, FCI and PSI. Feature selection is on average 25% (GDM-M) and 16% (GDM-F) more accurate than the best individual roll-off.

Table 7-5 DEPRESSION CLASSIFICATION ACCURACY (%), SENSITIVITY AND SPECIFICITY USING A 2-STAGE FILTER-WRAPPER FEATURE SELECTION (mRMR/SVM) FOR THE SPECTRAL ROLL-OFF PARAMETERS COMPARED TO THE ENTIRE FEATURE SET

<table>
<thead>
<tr>
<th>Features</th>
<th>Feature Selection</th>
<th>Entire Feature Set</th>
<th>Best Single Roll-Off</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sens</td>
<td>Spec</td>
<td>Acc</td>
</tr>
<tr>
<td>GDM-M</td>
<td>EPI</td>
<td>83.6</td>
<td>78.1</td>
</tr>
<tr>
<td></td>
<td>FCI</td>
<td>85.3</td>
<td>81.3</td>
</tr>
<tr>
<td></td>
<td>PSI</td>
<td>86.3</td>
<td>79.2</td>
</tr>
<tr>
<td>GDM-F</td>
<td>EPI</td>
<td>64.2</td>
<td>61.9</td>
</tr>
<tr>
<td></td>
<td>FCI</td>
<td>68.3</td>
<td>52.3</td>
</tr>
<tr>
<td></td>
<td>PSI</td>
<td>86.3</td>
<td>89.9</td>
</tr>
</tbody>
</table>
7.7.4 Effectiveness of Spectral and Prosodic Feature Sub-Categories (EXP4)

EXP4 has examined various prosodic and spectral features previously documented in depression detection [158][139][424][109] including ORI-DB studies [108][140][138][107], compared to the new range of spectral roll-offs determined from EXP3 in Section 7.7.3. A more effective state-of-the-art classifier is used with an optimized SVM system [406][408]. ORI-DB alterations, compared to past studies using the same dataset, use cleaned speech signals and only dyadic conversations.

Table 7-6 Depression Classification Accuracy, Sensitivity and Specificity using SVM with Prosodic (P) and Spectral (S) Features for Male Subjects (GDM-M) for Each Interaction (EPI, FCI, PSI)

<table>
<thead>
<tr>
<th>Features</th>
<th>Male Adolescents</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EPI</td>
</tr>
<tr>
<td></td>
<td>Sens</td>
</tr>
<tr>
<td>P</td>
<td></td>
</tr>
<tr>
<td>Fo</td>
<td>71.4</td>
</tr>
<tr>
<td>LogE</td>
<td>82.3</td>
</tr>
<tr>
<td>Jitter</td>
<td>68.1</td>
</tr>
<tr>
<td>Shimmer</td>
<td>62.4</td>
</tr>
<tr>
<td>S</td>
<td></td>
</tr>
<tr>
<td>Formants</td>
<td>70.9</td>
</tr>
<tr>
<td>PSD</td>
<td>78.9</td>
</tr>
<tr>
<td>Flux</td>
<td>56.9</td>
</tr>
<tr>
<td>Centroid</td>
<td>68.1</td>
</tr>
<tr>
<td>Entropy</td>
<td>65.2</td>
</tr>
<tr>
<td>Roll-off</td>
<td>83.6</td>
</tr>
</tbody>
</table>

Table 7-7 Depression Classification Accuracy, Sensitivity and Specificity using SVM with Prosodic (P) and Spectral (S) Features for Female Subjects (GDM-F) for Each Interaction (EPI, FCI, PSI)

<table>
<thead>
<tr>
<th>Features</th>
<th>Female Adolescents</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EPI</td>
</tr>
<tr>
<td></td>
<td>Sens</td>
</tr>
<tr>
<td>P</td>
<td></td>
</tr>
<tr>
<td>Fo</td>
<td>72.5</td>
</tr>
<tr>
<td>LogE</td>
<td>60.8</td>
</tr>
<tr>
<td>Jitter</td>
<td>66.1</td>
</tr>
<tr>
<td>Shimmer</td>
<td>65.8</td>
</tr>
<tr>
<td>S</td>
<td></td>
</tr>
<tr>
<td>Formants</td>
<td>58.8</td>
</tr>
<tr>
<td>PSD</td>
<td>75.9</td>
</tr>
<tr>
<td>Flux</td>
<td>72.2</td>
</tr>
<tr>
<td>Centroid</td>
<td>63.0</td>
</tr>
<tr>
<td>Entropy</td>
<td>57.1</td>
</tr>
<tr>
<td>Roll-off</td>
<td>64.2</td>
</tr>
</tbody>
</table>
Spectral and prosodic feature SVM classification performance, defined by accuracy, sensitivity and specificity, are given in Table 7-6 and Table 7-7 for GDM-M and GDM-F. Comparing features, both GDM setups, showed within the prosodic group LogE and F0 were generally the strongest discriminators and jitter and shimmer the worst, which is not surprising given jitter and shimmer had no statistical significance. Spectral features had the best result for spectral roll off range, slightly lower for formants and PSD, with flux, centroid and entropy are noticeably worse.

Results were consistent with psychological studies of speech as depression indicators, severity and treatment efficacy [437]. The best features were in line with clearly associated acoustic correlations of F0 [104][397][371][374][436][419][293], formants [156][437][432][109] power distribution (PSD) [390][156] and energy (LogE) [104] [376][457][437]. Additionally, coinciding with current literature the spectral features were more accurate than prosodic [158][139][108][140][424][109].

All GDM-M spectral features were more accurate than GDM-F, on average by 9% (EPI) and 11% (FCI). Similar, prosodic features detect male depression more accurately, except F0, by an average of 3%. In PSI, GDM-F had a higher accuracy for F0, jitter, PSD, entropy and roll-off, but on average GDM-M is 3% and 1% more accurate for prosodic and spectral features. There was considerable accuracy difference between genders for PSD, F0, LogE, formants and especially roll-off.

Comparing depression detection effectiveness in relation to interaction showed PSI had the best performance for most features. In GDM-F the average accuracy of prosodic features in PSI was 64% (7% higher than EPI and FCI) and spectral 68% (12% and 13% higher than EPI and FCI). Similar in GDM-M, to a lesser extent, PSI was the optimal interaction in both prosodic and spectral features with 61% (1% higher than EPI and FCI) and 68% (3% and 2% more than EPI and FCI).
7.7.5 Effectiveness of Prosodic and Spectral Feature Categories (EXP5)

The idea of EXP5 was to replicate previous documented studies, mainly from an adolescent depression approach using prosodic and spectral features [108][140][92], including the new range of spectral roll-offs determined from EXP3 in Section 7.7.3.

Similar to procedures in past studies, the acoustic features from EXP4 are fused into larger feature sets as Prosodic (P) and Spectral (S) categories, outlined in Table 7-1. Results are given in Table 7-8 for the P category and two implementations of the spectral category, using the set from past studies S* [140][105][92][108], and a new set, S, including a range of spectral roll-offs outlined in Section 7.7.3.

The results indicated the new spectral category (S), including the optimized roll-off parameters, was the best category. The original spectral category (S*), with only a single roll-off value was on average lower by 26% (GDM-M) and 22% (GDM-F).

Furthermore the P category was on average, across topics, 18% (GDM-M) and 25% (GDM-F) less accurate than S category. This agreed with EXP4 that determined individual spectral features were more accurate than prosodic features. This was consistent with depression studies indicating S is better than P [139][108][140][424].

GDM-M was more accurate than GDM-F, except for EPI/S* with GDM-F superior. On average, for three interactions, P, S and S* groups for GDM-M are 12%, 5% and 1% more accurate than GDM-F. Considering S* was missing the spectral roll-off this would suggest this feature is more helpful in male depression detection.

<table>
<thead>
<tr>
<th>Features</th>
<th>GDM-M</th>
<th>GDM-F</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sens</td>
<td>Spec</td>
</tr>
<tr>
<td>Prosodic (P)</td>
<td>83.8</td>
<td>71.9</td>
</tr>
<tr>
<td>Spectral (S)</td>
<td>99.3</td>
<td>95.4</td>
</tr>
<tr>
<td>Spectral* (S*)</td>
<td>80.5</td>
<td>61.2</td>
</tr>
<tr>
<td>Prosodic (P)</td>
<td>72.7</td>
<td>62.7</td>
</tr>
<tr>
<td>Spectral (S)</td>
<td>98.3</td>
<td>91.7</td>
</tr>
<tr>
<td>Spectral* (S*)</td>
<td>78.5</td>
<td>67.3</td>
</tr>
</tbody>
</table>

Table 7-8 Depression Classification Accuracy, Sensitivity and Specificity using SVM with Categorical Prosodic, Spectral* and Spectral (Roll-Off) Feature Sets for GDM-M and GDM-F and Each Interaction
It was observed that for GDM-M the P, S and S* groups were most or equal most accurate in the PSI, even though the accuracy difference between interactions was minimal. The GDM-F showed EPI was the best for S and S* and the FCI is the best for P, which is in contrast to the individual features in which PSI was optimal.

Table 7-9 Accuracy Difference of P, S*, and S Categories Compared to the Maximum, Minimum and Average of the Individual Sub-Categorical Features

<table>
<thead>
<tr>
<th>Features</th>
<th>EPI</th>
<th>FCI</th>
<th>PSI</th>
<th>Avg. Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Avg</td>
<td>Min</td>
<td>Max</td>
<td>Avg</td>
</tr>
<tr>
<td><strong>GDM-M</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prosodic (P)</td>
<td>17.6</td>
<td>23.6</td>
<td>13.3</td>
<td>20.0</td>
</tr>
<tr>
<td>Spectral (S)</td>
<td>32.0</td>
<td>42.1</td>
<td>16.5</td>
<td>31.6</td>
</tr>
<tr>
<td>Spectral* (S*)</td>
<td>5.6</td>
<td>15.7</td>
<td>-9.9</td>
<td>6.0</td>
</tr>
<tr>
<td><strong>GDM-F</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prosodic (P)</td>
<td>10.4</td>
<td>15.0</td>
<td>1.8</td>
<td>11.1</td>
</tr>
<tr>
<td>Spectral (S)</td>
<td>38.6</td>
<td>43.7</td>
<td>32.0</td>
<td>37.2</td>
</tr>
<tr>
<td>Spectral* (S*)</td>
<td>16.4</td>
<td>21.5</td>
<td>9.8</td>
<td>14.7</td>
</tr>
</tbody>
</table>

The individual prosodic and spectral features, given in Table 7-6 and Table 7-7 for GDM-M and GDM-F, did not all perform well independently and were improved when combined into S and P categories. Table 7-9 shows a comparison of the individual features and category given by the maximum, minimum and average accuracy difference for the individual features in the corresponding category (P or S).

Averaged for all topics (EPI, FCI, PSI) the P group has improved on the average individual features is 19% and 8% for the GDM-M and GDM-F. The S* and S category accuracies improves on average, compared to all features in the respective category by 5% and 31% for GDM-M and 11% and 33% for the GDM-F.

It is noted that fusing features improves upon all of the corresponding individual feature sets, with only a few exceptions. These include S* being lower than the spectral roll-off and to a less extent formants and PSD for all GDM-M (all topics) GDM-F (PSI). F0 is the only individual feature from the prosodic category that outperforms the entire P group (GDM-F FCI and PSI).
7.7.6 Cepstral (MFCC) and TEO-based (TEO-CB-Auto-Env) Features (EXP6)

Continuing with the acoustic feature categories (Table 7-1) identified in past studies [107][140][105]; both cepstral (MFCC) and TEO-based (TEO-CB-Auto-Env) feature sets are used as described in Sections 7.3.3 and 7.3.4. Table 7-10 gives the performance of both MFCC and TEO for each GDM and all interactions.

PSI was generally more accurate than the other two interactions, with only a single exception GDM-M/TEO/FCI combination. On average for both genders the average improvement of PSI is between 2%-4% for the MFCC and 0%-4% for TEO compared to FCI and PSI.

In most of the setups depression detection was easier to detect in the male adolescents compared to females. On average, across topics, MFCC and TEO are 10% and 5% more accurate for GDM-M compared to GDM-F. In all setups GDM-M was more accurate than GDM-F, except PSI/TEO combination with GDM-F <1% higher.

<table>
<thead>
<tr>
<th>Features</th>
<th>EPI</th>
<th>FCI</th>
<th>PSI</th>
</tr>
</thead>
<tbody>
<tr>
<td>GDM-M</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MFCC</td>
<td>72.6</td>
<td>74.0</td>
<td>76.6</td>
</tr>
<tr>
<td>TEO</td>
<td>86.1</td>
<td>70.6</td>
<td>84.2</td>
</tr>
<tr>
<td>GDM-F</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MFCC</td>
<td>66.1</td>
<td>66.4</td>
<td>69.0</td>
</tr>
<tr>
<td>TEO</td>
<td>67.4</td>
<td>70.1</td>
<td>75.3</td>
</tr>
</tbody>
</table>

MFCC have a slightly higher accuracy, on average 2% more accurate, compared to TEO in GDM-M case for all topics; although effectively comparable and in the GDM-F case TEO is best, compared to MFCC, by an average of 3%.
The GDM (GDM-M or GDM-F) with higher detection rates varies amongst feature/topic combinations. This suggests multiple features/dependencies determine the optimal gender setup. This is consistent with past studies having contradictory results comparing TEO and MFCC, some show MFCC are better in depression detection [107], and in contrast others have shown TEO performed best [108] [140].

It should be noted that the critical bands are derived in the other studies using Gabor filters and only use 15 bands, in comparison to the 17 bands in this thesis from the Wavelet Packet approach.

Additionally the speech dataset in this thesis is cleaned, which could effect on the overall result. This is based on evidence that in low noise MFCC are optimal and in noisy conditions TEOs outperform MFCC accuracy [189].

The cleaned signal in these experiments could explain why a similar study with a different subset of the same ORI-DB achieved a lower average accuracy, for GDM and topics, with MFCC (65.8% Low et al. [140], 69.6% EXP6). Yet, TEO was higher in the previous study (83% Low et al. [140], 70.4% EXP6), although it has to be noted in [140] more data was used and were subject-based.
7.7.7 Robustness of MFCC and TEO-CB-Auto-Env with AWGN (EXP7)

Studies suggest that TEO parameters are robust to additive noise and noisy environments [228][229][364] and have largely eliminated the effect of noise [364]. In contrast, MFCC are vulnerable with noisy speech and result in lower performance in many speech recognition tasks [189][363][427][524][525][526][527].

Past studies have investigated the noise robustness of MFCC and TEO in relation to speech processing, speaker characterization and emotion recognition using additive pink or white noise at multiple SNR levels ranging from 0dB to 20, 30 or 50dB generally in either 5dB or 10dB increments [524][525][526][527].

To investigate and further explain the previous results in EXP6 an examination has been conducted with varying levels of additive white Gaussian noise (AWGN) applied to speech with different SNR levels of 0dB, 10dB and 20dB.

SVM depression classification performance using MFCC and TEO extracted from speech signal with various levels of AWGN is given in Table 7-11 and Table 7-12 for each interaction (EPI, FCI, PSI) and Figure 7-9 averaged for all interactions. The AWGN has similar observations to noise free features such that GDM-M is more accurate than GDM-F and mostly PSI and FCI are better than EPI.

![Comparison of AWGN SNR Levels](image)

Figure 7-9 Average accuracy (%) of the three interactions comparing the performance of the AWGN SNR levels for the MFCC and TEO features and each GDM
Table 7-11 DEPRESSION CLASSIFICATION ACCURACY (%), SENSITIVITY AND SPECIFICITY USING SVM WITH MFCC EXTRACTED FROM SPEECH SIGNALS WITH VARIOUS LEVELS OF AWG NOISE FOR GDM-F AND GDM-M AND EACH INTERACTION

<table>
<thead>
<tr>
<th>MFCC with AWGN SVM classification Performance</th>
<th>EPI</th>
<th>FCI</th>
<th>PSI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Features</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GDM-M</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Noise</td>
<td>72.6</td>
<td>73.5</td>
<td>73.0</td>
</tr>
<tr>
<td>20dB</td>
<td>71.6</td>
<td>69.6</td>
<td>70.6</td>
</tr>
<tr>
<td>10dB</td>
<td>69.3</td>
<td>63.1</td>
<td>66.2</td>
</tr>
<tr>
<td>0dB</td>
<td>59.6</td>
<td>57.4</td>
<td>58.5</td>
</tr>
<tr>
<td>GDM-F</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Noise</td>
<td>66.1</td>
<td>65.6</td>
<td>65.8</td>
</tr>
<tr>
<td>20dB</td>
<td>64.8</td>
<td>64.2</td>
<td>64.5</td>
</tr>
<tr>
<td>10dB</td>
<td>63.2</td>
<td>59.1</td>
<td>61.1</td>
</tr>
<tr>
<td>0dB</td>
<td>57.3</td>
<td>56.2</td>
<td>56.8</td>
</tr>
</tbody>
</table>

Table 7-12 DEPRESSION CLASSIFICATION ACCURACY (%), SENSITIVITY AND SPECIFICITY USING SVM WITH TEO EXTRACTED FROM SPEECH SIGNALS WITH VARIOUS LEVELS OF AWG NOISE FOR GDM-F AND GDM-M AND EACH INTERACTION

<table>
<thead>
<tr>
<th>TEO with AWGN SVM classification Performance</th>
<th>EPI</th>
<th>FCI</th>
<th>PSI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Features</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GDM-M</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Noise</td>
<td>86.1</td>
<td>59.4</td>
<td>72.5</td>
</tr>
<tr>
<td>20dB</td>
<td>72.4</td>
<td>72.3</td>
<td>72.4</td>
</tr>
<tr>
<td>10dB</td>
<td>64.8</td>
<td>77.9</td>
<td>71.3</td>
</tr>
<tr>
<td>0dB</td>
<td>66.5</td>
<td>67.6</td>
<td>67.0</td>
</tr>
<tr>
<td>GDM-F</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Noise</td>
<td>67.4</td>
<td>58.1</td>
<td>62.8</td>
</tr>
<tr>
<td>20dB</td>
<td>59.4</td>
<td>63.6</td>
<td>61.5</td>
</tr>
<tr>
<td>10dB</td>
<td>65.0</td>
<td>57.1</td>
<td>61.0</td>
</tr>
<tr>
<td>0dB</td>
<td>65.9</td>
<td>56.1</td>
<td>60.9</td>
</tr>
</tbody>
</table>

In the GDM-M and GDM-F/EPI cases the cleaned (noise-free) signal, from EXP6, MFCC outperformed TEO. In contrast with noise TEO was generally better in every GDM and topic. MFCC performance degraded more substantially compared to TEO as the noise increased. In fact TEO performance comparatively had minimal change with AWGN. This suggested TEO was robust to noise, which concurs with current literature investigating the effect of noise [228][229][189][363][364][427].
7.7.8 Glottal Category (G) time (GTD) and frequency (GFD) features (EXP8)

The glottal waveform plays an important role in speech production and is greatly affected by emotional/stressful speech [401]. Glottal features have a strong correlation, in depression and provide excellent depression detection rates [105][140][108], predominantly outperforming other features [92][106][65].

As suggested by previous depression studies [105][140][106][65][92][108], EXP8 utilizes the glottal category (G) containing nine glottal time-domain (GTD) and three glottal frequency-domain (GFD) parameters. Further analysis was conducted by separating into individual GTD and GFD subcategories. Table 7-13 gives the results of depression classification using G category and GTD and GFD subcategories.

The GTD is more efficient than GFD parameters across all interactions by an average of 10% for GDM-M and 7.5% for GDM-F. Fusing the time and frequency features (GTD+GFD) enhances the accuracy for each of the domains in all setups. The GTD accuracy is on average increased by 6.3% and 7.9% with the inclusion of GFD for GDM-M and GDM-F. Similar, the inclusion of GTD improves on GFD by 16.2% and 15.4% for GDM-M and GDM-F. This suggests both GTD and GFD are important depression discriminators and contain complementary information.

### Table 7-13 Depression Classification Accuracy (%), Sensitivity and Specificity Using SVM with Glottal Time-Domain (GTD) and Frequency-Domain (GFD) Features for GDM-M and GDM-F and Each Interaction

<table>
<thead>
<tr>
<th>Features</th>
<th>EPI</th>
<th>FCI</th>
<th>PSI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sens</td>
<td>Spec</td>
<td>Acc</td>
</tr>
<tr>
<td>GDM-M</td>
<td>GTD</td>
<td>74.1</td>
<td>73.8</td>
</tr>
<tr>
<td></td>
<td>GFD</td>
<td>77.6</td>
<td>47.0</td>
</tr>
<tr>
<td></td>
<td>GT+GF</td>
<td>86.4</td>
<td>81.5</td>
</tr>
<tr>
<td>GDM-F</td>
<td>GTD</td>
<td>71.0</td>
<td>74.9</td>
</tr>
<tr>
<td></td>
<td>GFD</td>
<td>64.8</td>
<td>67.4</td>
</tr>
<tr>
<td></td>
<td>GT+GF</td>
<td>81.1</td>
<td>82.0</td>
</tr>
</tbody>
</table>

The results showed glottal features were more important in detecting depression in males compared to females. This was based on the increased accuracy of GDM-M for all features and interactions compared to GDM-F, with one exception (GFD/EPI).
7.7.9 Glottal Waveform (GW) features (G-MFCC and G-TEO) (EXP9)

The importance of the glottal category (G) was investigated by deriving MFCC and TEO from the glottal waveform (GW), outlined in Section 7.3.5.3. Glottal waveform TEO and MFCC have been used in emotion recognition [238][342][367] including research from within this dissertation (see Chapter Four and published in [497]).

The glottal waveform TEO (G-TEO) has already been used in depression classification, improving upon TEO extracted from the speech signal [65][147]. The glottal waveform MFCC (G-MFCC), to our knowledge has not yet been used in a depression detection application. Both could be beneficial, considering depression is an emotion regulation disorder [24] associated with emotional disturbances [449].

Table 7-14 supplies the performance (accuracy, sensitivity and specificity) of the glottal waveform features (G-MFCC and G-TEO) for each GDM and topic. A comparison of GDM showed G-TEO detects depression easier in GDM-M by an average of 2.6%. Similar, G-MFCC was more accurate for GDM-M, with the exception of PSI, by an average of 0.4%. In each GDM both G-MFCC and G-TEO have the highest accuracy in either the FCI or PSI.

In all set-ups G-TEO had higher accuracy than G-MFCC on average by 4.2% and 6.4% for both GDM-M and GDM-F. In contrast the respective speech waveform parameters, EXP6 in Section 7.7.6, showed MFCC was slightly better. This indicates the glottal waveform was more helpful in depression classification using TEO.

<table>
<thead>
<tr>
<th>Features</th>
<th>EPI</th>
<th>FCI</th>
<th>PSI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sens</td>
<td>Spec</td>
<td>Acc</td>
</tr>
<tr>
<td>GDM-M</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G-MFCC</td>
<td>72.1</td>
<td>64.3</td>
<td>68.1</td>
</tr>
<tr>
<td>G-TEO</td>
<td>73.6</td>
<td>77.0</td>
<td>75.3</td>
</tr>
<tr>
<td>GDM-F</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G-MFCC</td>
<td>76.5</td>
<td>57.2</td>
<td>66.7</td>
</tr>
<tr>
<td>G-TEO</td>
<td>79.2</td>
<td>67.2</td>
<td>73.2</td>
</tr>
</tbody>
</table>
The average accuracy for MFCC and TEO derived from glottal waveform and speech waveform are compared in Figure 7-10. G-TEO improved on TEO by an average of 2.7% (GDM-M) and 7.1% (GDM-F). On average G -MFCC was 4% higher than MFCC (GDM-F). G-MFCC/GDM-M was the only case the glottal waveform feature did not improve on the speech feature (MFCC), with 3% average reduction.

7.7.10 Combine glottal category (G) with categories (S, P, C, TEO, GW) (EXP10)

Depression related studies have indicated the addition of glottal features enhances classification compared to stand alone features (S, P, TEO, MFCC)[105][140][106]. As suggested by these previous studies, EXP10 investigated the glottal category (G) from Section 7.7.8 (EXP8) (GTD and GFD parameters) fused with P and S categories from Section 7.7.5 (EXP5) and Cepstral (MFCC) and TEO from Section 7.7.6 (EXP6). Based on the importance of the glottal waveform a combination of G and the glottal waveform category (GW), G-MFCC and G-TEO from Section 7.7.9 (EXP9), was proposed.
Table 7-15 Depressed Classification Accuracy (%), Sensitivity and Specificity using SVM with Prosodic (P), Spectral (S), MFCC, TEO, G-MFCC and G-TEO Category Features in Addition to the Glottal Category Features (+G) for Male Subjects (GDM-M) for Each Interaction

<table>
<thead>
<tr>
<th>Features</th>
<th>Male Adolescents (GDM-M) Feature Combined with “+G”</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCI</td>
<td>EPI</td>
</tr>
<tr>
<td>S</td>
<td>+G</td>
</tr>
<tr>
<td>P</td>
<td>+G</td>
</tr>
<tr>
<td>MFCC</td>
<td>+G</td>
</tr>
<tr>
<td>G-MFCC</td>
<td>+G</td>
</tr>
<tr>
<td>TEO</td>
<td>+G</td>
</tr>
<tr>
<td>G-TEO</td>
<td>+G</td>
</tr>
</tbody>
</table>

Table 7-16 Depressed Classification Accuracy (%), Sensitivity and Specificity using SVM with Prosodic (P), Spectral (S), MFCC, TEO, G-MFCC and G-TEO Category Features in Addition to the Glottal Category Features (+G) for Male Subjects (GDM-M) for Each Interaction

<table>
<thead>
<tr>
<th>Features</th>
<th>Female Adolescents (GDM-F) Feature Combined with “+G”</th>
</tr>
</thead>
<tbody>
<tr>
<td>EPI</td>
<td>FCI</td>
</tr>
<tr>
<td>S</td>
<td>+G</td>
</tr>
<tr>
<td>P</td>
<td>+G</td>
</tr>
<tr>
<td>MFCC</td>
<td>+G</td>
</tr>
<tr>
<td>G-MFCC</td>
<td>+G</td>
</tr>
<tr>
<td>TEO</td>
<td>+G</td>
</tr>
<tr>
<td>G-TEO</td>
<td>+G</td>
</tr>
</tbody>
</table>

Table 7-15 and Table 7-16 provide classification performance of each feature category (P, S, C, T, GW) combined with G. PSI had the highest accuracy in every case, except two instances (GDM-M/P+G and GDM-F/S+G). There was an even spread throughout features and topics as to which GDM was superior. For the most part GDM-F was better than GDM with the S+G and P+G sets. GDM-M was better than the GDM-F in G-TEO+G and TEO+G sets.

Figure 7-11 Average accuracy (%) of adolescent depression classification, for all interactions, comparing the “+G” parameters for the GDM-M and GDM-F.
The average accuracy of the topics is illustrated in Figure 7-11 comparing the performance of the “+G” feature combinations. “+G” features reveal similar patterns to the previous experiments, such that MFCC+G was the worst and S+G the best set. There was consistency with EXP6 and EXP9 as MFCC+G was improved using the glottal waveform (G-MFCC+G) by an average of 2% for GDM-M and GDM-F. G-TEO+G was improved by 1% compared to TEO+G for GDM-M and GDM-F. As with EXP5, P+G was notably better than S+G by 9% (GDM-M) and 11% (GDM-F).

Table 7-17 gives the difference in classification accuracy between a given feature with “+G” and without. The results revealed the addition of G improved all features (S, P, MFCC, TEO, G-MFCC and G-TEO) for all topics and GDMs. The only exception is the S category in which “+G” lowered the performance for the GDM-M.

In general the addition of the G category improved accuracy more for GDM-F, which could indicate glottal importance in females. In fact “+G” features have an even spread to which gender is more accurate, where as the features alone mostly GDM-M was better. It can noted that for each “+G” feature set outperformed independent G for every feature except for MFCC+G and G-MFCC+G.
7.7.11 Combine Spectral category (S) with categories (P, C, TEO, G, GW) (EXP11)

Previous results signified the S category, from Section 7.7.5 (EXP5), has the best performance, with over 90% accuracy, and could supply supplementary information fused with the other categories to increase performance. The depression classification performance using P, MFCC, TEO, G, G-MFCC and G-TEO fused with “+S” is supplied in Table 7-18 and Table 7-19 for GDM-M and GDM-F respectively.

Table 7-18 Depression Classification Accuracy (%), Sensitivity and Specificity using SVM with Prosodic (P), Glottal (G), MFCC, TEO, G-MFCC and G-TEO category features in addition to the Spectral category features (+S) for male subjects (GDM-M) for each interaction

<table>
<thead>
<tr>
<th>Features</th>
<th>Male Adolescents (GDM-M) Feature Combined with “+S”</th>
<th>EPI</th>
<th>FCI</th>
<th>PSI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sens</td>
<td>Spec</td>
<td>Acc</td>
<td>Sens</td>
</tr>
<tr>
<td>G</td>
<td>+S</td>
<td>1.3</td>
<td>90.5</td>
<td>93.7</td>
</tr>
<tr>
<td>P</td>
<td>+S</td>
<td>99.7</td>
<td>98.2</td>
<td>98.9</td>
</tr>
<tr>
<td>MFCC</td>
<td>+S</td>
<td>95.9</td>
<td>93.7</td>
<td>94.8</td>
</tr>
<tr>
<td>G-MFCC</td>
<td>+S</td>
<td>98.3</td>
<td>94.7</td>
<td>96.5</td>
</tr>
<tr>
<td>TEO</td>
<td>+S</td>
<td>98.6</td>
<td>90.8</td>
<td>94.8</td>
</tr>
<tr>
<td>G-TEO</td>
<td>+S</td>
<td>97.7</td>
<td>95.5</td>
<td>96.6</td>
</tr>
</tbody>
</table>

Table 7-19 Depression Classification Accuracy (%), Sensitivity and Specificity using SVM with Prosodic (P), Glottal (G), MFCC, TEO, G-MFCC and G-TEO category features in addition to the Spectral category features (+S) for female subjects (GDM-F) for each interaction

<table>
<thead>
<tr>
<th>Features</th>
<th>Female Adolescents (GDM-F) Feature Combined with “+S”</th>
<th>EPI</th>
<th>FCI</th>
<th>PSI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sens</td>
<td>Spec</td>
<td>Acc</td>
<td>Sens</td>
</tr>
<tr>
<td>G</td>
<td>+S</td>
<td>99.5</td>
<td>96.2</td>
<td>97.9</td>
</tr>
<tr>
<td>P</td>
<td>+S</td>
<td>98.4</td>
<td>89.9</td>
<td>94.1</td>
</tr>
<tr>
<td>MFCC</td>
<td>+S</td>
<td>63.4</td>
<td>99.2</td>
<td>81.4</td>
</tr>
<tr>
<td>G-MFCC</td>
<td>+S</td>
<td>94.8</td>
<td>94.1</td>
<td>94.5</td>
</tr>
<tr>
<td>TEO</td>
<td>+S</td>
<td>94.0</td>
<td>92.9</td>
<td>93.5</td>
</tr>
<tr>
<td>G-TEO</td>
<td>+S</td>
<td>97.5</td>
<td>92.0</td>
<td>94.7</td>
</tr>
</tbody>
</table>

Consistent with previous experiments PSI was the most accurate topic with only three exceptions, G-MFCC+S/GDM-M and TEO+S/GDM-M with FCI and G+S/GDM-F with EPI. GDM-M has higher accuracy than GDM-F for all features except G+S; indicating the importance of glottal features in females. Moreover, S+G is the top GDM-F feature and P+S is highest performing GDM-M feature.
Figure 7-12 gives the average depression classification accuracy comparing the performance of the “+S” feature sets and was consistent with EXP6 and EXP9.

For example G-MFCC+S for GDM-M and GDM-F are on average 2% and 5% better than MFCC+S similar G-TEO+S is 1% and 2% better than TEO+S.

Table 7-20 gives the accuracy difference with the inclusion of “+S” compared to the individual feature (i.e. without “+S”) from previous sections. The results suggest “+S” improved all sets (G, P, MFCC, TEO, G-MFCC and G-TEO) for every case. It can be noted that in the GDM-F the “+S” feature sets were mostly better than S independently but not true in the GDM-M case. “+S” increased the accuracy by more than “+G” set, which was expected as S performed better independently.
7.7.12 Depression detection in adolescents using parents speech (EXP12)

There can be difficulties in gathering speech data from children including possible ethics issues or adolescents not willing to participate. Another problem is the issue of voice changing during adolescence affecting the detection system. Considering these issues it could be useful to examine parent’s speech to determine if it is possible to detect adolescent depression. This can be justified as family relationships are correlated to adolescent depression [136][134][137][24]. For an initial investigation the previous experiments using the six main feature categories as follows:

- Spectral (S): EXP5 from Section 7.7.5 formant, PSD, flux, centroid, entropy, roll-off
- Prosodic (P): EXP5 from Section 7.7.5 using F0, LogE, jitter and shimmer
- Cepstral (C): EXP6 from Section 7.7.6 using MFCC
- TEO (T): EXP6 from Section 7.7.6 using TEO-CB-Auto-Env
- Glottal (G): EXP8 from Section 7.7.8 with GTD combined with GFD
- Glottal Waveform (GW): EXP9 from Section 7.7.9 with G-MFCC and G-TEO

Table 7-21 gives SVM adolescent depression detection performance from parent speech, either a father (GDM-M) or mother (GDM-F), comparing seven feature sets. The results revealed PSI gave the highest accuracy for all features in GDM-M and all except P and G-MFCC in GDM-F Another similar observation was that for the majority of features adolescent depression was easier to detect from speech of mothers (i.e. GDM-F) with only a few exceptions.

Glottal waveform MFCC (G-MFCC) improved on the speech waveform MFCC by an average of 6.3% and 2.8% in GDM-F and GDM-M respectively. Furthermore G-TEO improved upon TEO by an average of 6.6% (GDM-F) and 3.6% (GDM-M). Comparing all features showed for both GDM-M and GDM-F the worst for all interactions was MFCC on average 54%.
The GDM-M case the S category was clearly the optimal feature set in all interactions on average 68%. In GDM-F the top feature ranges over the topics although G-TEO has the highest average of 68%. The overall best GDM-F result was 69% in G-TEO/FCI case and 71% for GDM-M with S/PSI case.

Table 7-21 Adolescent Depression Classification Accuracy, Sensitivity and Specificity Using SVM with Acoustic Feature Categories Prosodic (P), Spectral (S), Glottal (G), Cepstral (C), TEO (T) and Glottal Waveform (GW) Extracted from Their Parent’s Speech for Fathers (GDM-M) and Mothers (GDM-F) and Each Interaction

<table>
<thead>
<tr>
<th>Features</th>
<th>GDM-M</th>
<th>GDM-F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prosodic (P)</td>
<td>Sens 55.5</td>
<td>Sens 73.3</td>
</tr>
<tr>
<td>Spectral (S)</td>
<td>Sens 68.5</td>
<td>Sens 58.1</td>
</tr>
<tr>
<td>MFCC (C)</td>
<td>Sens 54.1</td>
<td>Sens 64.1</td>
</tr>
<tr>
<td>TEO (T)</td>
<td>Sens 76.8</td>
<td>Sens 63.5</td>
</tr>
<tr>
<td>GT+GF (G)</td>
<td>Sens 63.5</td>
<td>Sens 56.7</td>
</tr>
<tr>
<td>G-MFCC (GW)</td>
<td>Sens 56.7</td>
<td>Sens 56.1</td>
</tr>
<tr>
<td>G-TEO (GW)</td>
<td>Sens 60.</td>
<td>Sens 59.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Features</th>
<th>EPI</th>
<th>PSI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sens</td>
<td>Spec</td>
<td>Acc</td>
</tr>
<tr>
<td>-----------</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Prosodic (P)</td>
<td>63.5</td>
<td>64.1</td>
</tr>
<tr>
<td>Spectral (S)</td>
<td>61.8</td>
<td>62.3</td>
</tr>
<tr>
<td>MFCC (C)</td>
<td>53.7</td>
<td>55.3</td>
</tr>
<tr>
<td>TEO (T)</td>
<td>56.3</td>
<td>57.8</td>
</tr>
<tr>
<td>GT+GF (G)</td>
<td>58.9</td>
<td>61.1</td>
</tr>
<tr>
<td>G-MFCC (GW)</td>
<td>59.7</td>
<td>60.5</td>
</tr>
<tr>
<td>G-TEO (GW)</td>
<td>59.7</td>
<td>60.5</td>
</tr>
</tbody>
</table>

Figure 7-13 gives the average adolescent depression classification accuracy for the feature categories with a direct comparison of the parent and adolescent speech features. It was shown for all feature categories the adolescent speech was significantly more accurate than from the parent’s speech. This is intuitive; as it would be expected detection from the actual subject compared to a partner is easier.

The adolescent’s speech compared to parent speech for GDM-M is more accurate by an average of 29%, 23%, 17%, 14%, 15%, 14% and 20% for S, G, P, G-TEO, TEO, G-TEO, G-MFCC and MFCC. Similarly the GDM-F case adolescent depression detection was easier from the actual adolescents compared to their parent by an average of 27%, 17%, 7%, 7%, 7%, 8% and 11% for each feature respectively.
The results of adolescent depression detection from their parent’s speech could be based on parents adapting to depressive (non-depressed) children’s speech, which is consistent with psychological studies explaining that parents attune to emotional behavior of children [116][88][87].

Although external factors such as the parents depressive status is unknown and if they are depressed it could alter modeling. It is known that depression is more likely to suffer from depression [337], which is especially true in the case of depression in mothers [7]. This could possibly explain why the rates are generally higher in GDM-F parent case compared to GDM-M case.

It would be necessary to carry on from this preliminary experiment with a purpose-met database, ideally for adolescent depression detection it would be best to know the parent’s depressive status. This way multiple models could be generated to detect adolescent depression from parent depression-status dependent models.
7.8 Discussion and Conclusions

This research has investigated gender dependence, importance of conversational topic and compared features from six categories including new depression detection features: G-MFCC and mRMR optimized spectral roll-offs. It has been determined that adolescent depression can be detected in from both adolescent and parent speech.

A) Gender

In past studies there has been conflicting results as to which gender is easiest to detect depression. Some studies have found that females have highest depression detection rates \[105][138][107] and other studies males \[140][108]. In fact even within studies there is variation amongst features as to which gender is easiest to classify \[140][108][105][107].

This research confirmed the importance of gender in depression detection, proven by \textit{EXP2}, with an accuracy improvement with GDMs compared to GIM using \textit{F0}. The best GDM (GDM-M or GDM-F), throughout \textit{EXP3-EXP9}, varied amongst features and interactions. Most cases had better detection rates in males (GDM-M) and the only exceptions of GDM-F with higher accuracy involved the G category.

B) Comparing topics

The three interaction tasks that were designed to create different types of interactional contexts, chosen specifically as they have been found to elicit differential levels of happy, angry, and dysphoric affect \[96][97].

Success in all topics gives reason to believe many conversation types could be useful. For the majority of features and combinations PSI performed the best (\textit{EXP2-EXP10}) and EPI the least effective. This is consistent with previous studies that showed PSI was most effective in depression classification \[140][454][107].
This could be due to PSI being setup to evoke conflicting behavior, which is strongly correlated to depression in family interactions, and produce more noticeable speech changes [276]. Another study confirmed that differences in expressing positive and negative emotions effect acoustic depression detection[435].

C) Features

A summary all acoustic features average accuracies are ranked in Table 7-22 for GDM-M and GDM-F. The blue shaded cells represent acoustic feature derived from adolescent’s parents’ features and grey cells are adolescent features. The various grey shades separate the sub-category, category, “+G” and “+S” feature combinations.

The top 8 ranking adolescent depression detection acoustic features, for both GDM-M and GDM-F, contain the S category that incorporate the proposed optimized spectral roll-off range. The sub-category feature sets are generally improved when combined into feature categories and further enhanced with combinations of “+G” and more so “+S”. The parent’s feature sets are mostly on the low end and each feature is lower than the corresponding adolescent feature. More detailed discussion and summary on each of the types of features are explained in the following sections.
Table 7-22 RANKING OF AVERAGE ACCURACY (%) DEPRESSION CLASSIFICATION PERFORMANCE OF EACH TOPIC (EPI, FCI, PSI) AND GENDER (GDM-M AND GDM-F) USING THE ORI DATABASE SPEECH FOR THE ACOUSTIC FEATURE APPROACH. THE SHADE OF THE GREY CELLS SIGNIFY, FROM LIGHTEST TO DARKEST: SUB-CATEGORY, CATEGORY, “+G” AND “+S”. THE BLUE SHADE REPRESENTS THE PARENT FEATURES.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Feature</th>
<th>GDM-M</th>
<th>Feature</th>
<th>GDM-F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>P+S</td>
<td>98.83</td>
<td>G+S</td>
<td>97.13</td>
</tr>
<tr>
<td>2</td>
<td>Spectral (S)</td>
<td>97.50</td>
<td>S+G</td>
<td>97.13</td>
</tr>
<tr>
<td>3</td>
<td>G-TEO+S</td>
<td>97.00</td>
<td>P+S</td>
<td>96.73</td>
</tr>
<tr>
<td>4</td>
<td>G-MFCC+S</td>
<td>96.53</td>
<td>G-TEO+S</td>
<td>95.17</td>
</tr>
<tr>
<td>5</td>
<td>TEO+S</td>
<td>95.87</td>
<td>G-MFCC+S</td>
<td>94.60</td>
</tr>
<tr>
<td>6</td>
<td>MFCC+S</td>
<td>95.03</td>
<td>TEO+S</td>
<td>93.20</td>
</tr>
<tr>
<td>7</td>
<td>G+S</td>
<td>94.23</td>
<td>Spectral (S)</td>
<td>92.30</td>
</tr>
<tr>
<td>8</td>
<td>S+G</td>
<td>94.23</td>
<td>MFCC+S</td>
<td>89.23</td>
</tr>
<tr>
<td>9</td>
<td>G-TEO+G</td>
<td>90.27</td>
<td>G-TEO+G</td>
<td>88.67</td>
</tr>
<tr>
<td>10</td>
<td>TEO+G</td>
<td>89.33</td>
<td>TEO+G</td>
<td>87.60</td>
</tr>
<tr>
<td>11</td>
<td>P+G</td>
<td>85.60</td>
<td>P+G</td>
<td>86.40</td>
</tr>
<tr>
<td>12</td>
<td>Glottal GT+GF (G)</td>
<td>83.03</td>
<td>G-MFCC+G</td>
<td>81.73</td>
</tr>
<tr>
<td>13</td>
<td>Roll-off</td>
<td>82.30</td>
<td>Glottal GT+GF (G)</td>
<td>81.37</td>
</tr>
<tr>
<td>14</td>
<td>G-MFCC+G</td>
<td>82.07</td>
<td>MFCC+G</td>
<td>79.40</td>
</tr>
<tr>
<td>15</td>
<td>MFCC+G</td>
<td>80.03</td>
<td>G-TEO</td>
<td>75.07</td>
</tr>
<tr>
<td>16</td>
<td>Prosodic (P)</td>
<td>79.30</td>
<td>GTD</td>
<td>73.43</td>
</tr>
<tr>
<td>17</td>
<td>GTD</td>
<td>76.73</td>
<td>Fo</td>
<td>71.77</td>
</tr>
<tr>
<td>18</td>
<td>G-TEO</td>
<td>75.43</td>
<td>Spectral* (S*)</td>
<td>70.63</td>
</tr>
<tr>
<td>19</td>
<td>MFCC</td>
<td>74.40</td>
<td>Roll-off</td>
<td>70.50</td>
</tr>
<tr>
<td>20</td>
<td>PSD</td>
<td>74.03</td>
<td>G-MFCC</td>
<td>68.70</td>
</tr>
<tr>
<td>21</td>
<td>TEO</td>
<td>72.77</td>
<td>TEO</td>
<td>67.93</td>
</tr>
<tr>
<td>22</td>
<td>Formants</td>
<td>72.60</td>
<td>G-TEO Parent</td>
<td>67.60</td>
</tr>
<tr>
<td>23</td>
<td>Spectral* (S*)</td>
<td>71.43</td>
<td>Prosodic (P)</td>
<td>67.07</td>
</tr>
<tr>
<td>24</td>
<td>G-MFCC</td>
<td>71.27</td>
<td>Prosodic (P) Parent</td>
<td>66.28</td>
</tr>
<tr>
<td>25</td>
<td>Spectral (S) Parent</td>
<td>68.07</td>
<td>GFD</td>
<td>65.97</td>
</tr>
<tr>
<td>26</td>
<td>GFD</td>
<td>66.83</td>
<td>PSD</td>
<td>65.77</td>
</tr>
<tr>
<td>27</td>
<td>LogE</td>
<td>65.80</td>
<td>MFCC</td>
<td>64.77</td>
</tr>
<tr>
<td>28</td>
<td>F0</td>
<td>63.30</td>
<td>Spectral (S) Parent</td>
<td>64.73</td>
</tr>
<tr>
<td>29</td>
<td>Prosodic (P) Parent</td>
<td>62.83</td>
<td>Glottal GT+GF (G) Parent</td>
<td>64.43</td>
</tr>
<tr>
<td>30</td>
<td>G-TEO Parent</td>
<td>61.43</td>
<td>Formants</td>
<td>62.37</td>
</tr>
<tr>
<td>31</td>
<td>Glottal GT+GF (G) Parent</td>
<td>59.73</td>
<td>TEO Parent</td>
<td>60.93</td>
</tr>
<tr>
<td>32</td>
<td>Jitter</td>
<td>58.73</td>
<td>G-MFCC Parent</td>
<td>60.43</td>
</tr>
<tr>
<td>33</td>
<td>TEO Parent</td>
<td>57.87</td>
<td>Jitter</td>
<td>58.70</td>
</tr>
<tr>
<td>34</td>
<td>Entropy</td>
<td>57.10</td>
<td>Entropy</td>
<td>54.97</td>
</tr>
<tr>
<td>35</td>
<td>G-MFCC Parent</td>
<td>57.07</td>
<td>LogE</td>
<td>54.97</td>
</tr>
<tr>
<td>36</td>
<td>Centroid</td>
<td>56.53</td>
<td>MFCC Parent</td>
<td>54.10</td>
</tr>
<tr>
<td>37</td>
<td>Flux</td>
<td>54.70</td>
<td>Flux</td>
<td>52.97</td>
</tr>
<tr>
<td>38</td>
<td>MFCC Parent</td>
<td>54.27</td>
<td>Shimmer</td>
<td>52.20</td>
</tr>
<tr>
<td>39</td>
<td>Shimmer</td>
<td>53.70</td>
<td>Centroid</td>
<td>51.93</td>
</tr>
</tbody>
</table>
i. Sub categorical Features

On average the best sub-categorical P feature set (P: \(F_0\), LogE, jitter, shimmer) is LogE for GDM-M with an average of 65.8\% and \(F_0\) for GDM-F with 71.8\%. The best glottal subcategory (G: GTD and GFD) for both GDM-M and GDM-F is GTD with 76.7\% and 73.4\% respectively. Comparing glottal waveform sub-categorical features (GW: G-MFCC and G-TEO) found G-TEO was best at 75\% for both GDMs.

The shape of the glottal pulse can measure emotion from tension applied of vocal folds during opening and closing phases of quasi-periodic vibration [423], where as this may not be detected by \(F_0\), jitter or shimmer measurements, explaining the relatively better performance of G compared to P sub-categorical features.

A main finding was the improvement of depression detection using a range of spectral rolls and further improved with 2-stage mRMR/SVM optimized feature selection. This was the highest performing individual feature set from the S category (S: Formants, PSD, Flux, Centroid, Entropy, Roll-off) and all other categories (G, P, GW, C, T) with an average accuracy of 82.2\% and 70.5\% for GDM-M and GDM-F.

In addition it was found that TEO are robust to AWG noise with little accuracy variation and in contrast MFCC accuracy reduce with noise. This agreed with studies that showed MFCC are not robust to noise in recognition [189][363][427].

ii. Glottal Waveform Features vs. Speech Waveform Features

An interesting finding was the increased in accuracy when deriving TEO from the glottal waveform (G-TEO), consistent with previous studies [65][147]. The new glottal waveform MFCC (G-MFCC) improved on the speech waveform MFCC only in the GDM-F case. These results support suggestion on the important role of glottal waveforms on clinical depression and the effectiveness in depression recognition [140][92][106][65][105].
Studies have suggested that psychological difficulties of depressed persons shows physical changes, within vocal fold and tract, during speech production so that patterns of the glottal waveform reflect cues for speech depression classification[462][463]. This could explain why the glottal waveform features are more effective in detecting depressed subjects compared to speech waveform.

The number of supra-glottal vortices is related to emotional stress and hence depression [55]. Depressed speech has differing laminar and vortices air flow patterns compared to controls. This is captured by TEO and could provide a reason why glottal TEO was more effective than TEO and not evident in G-MFCC compared to MFCC.

### iii. Categorical Features

A comparison of the six categories using combined subcategories found the S category outperformed P, C, T, G and GW with an average of 97.5% for GDM-M and 92.3% for GDM-F. Overall the best individual result was attained for GDM-M/PSI case with 97.9% accuracy and 99.5%/5% sensitivity/specificity ratio. The results agreed with past studies that S is better than P [105][158][139][108][140][424][109].

### iv. Combination of feature categories

An examination of feature category combinations was conducted using the suggestion by previous successful studies by combining G category with the remaining categories (i.e. P+G, P+S,P+S+G) [140][105]. This was compared to using the top performing feature category S found from the experiments within this dissertation.

The addition of “+G” and “+S” both improved the performance of the stand-alone feature sets (P, C, T, GW and G or S). Overall the “+S” improved the performance of the stand-alone categories by more than the “+G” which would be expected given the enhanced results of S compared to G.
On average the highest result of “+S” occurred using “S+G” combined feature with an accuracy of 97.1% for the GDM-F. The best average feature combination in the GDM-M case was “S+P” with 98.8%. Overall the top result was attained with “S+P” GDM-M case for PSI reaching a top accuracy of 98.9% accuracy with sensitivity specificity ratios of 99.9%/98.9.

D) Parent
The results for detecting children’s depression from parent speech are promising, the best result obtained was the S group with the GDM-M reaching an average of 68% accuracy and the G-TEO for the GDM-F with an average of 68%. This could be seen as parents changing their speech patterns to contend with a depressed child.

A limitation of the preliminary results was the extraneous variable of the unknown mental status of the parent. Therefore, the system could actually have detected the parent’s mental status and not from their child.

This could be the case as those that are being treated for depression have a higher risk of their children developing depression during adolescence [337]. The ORI dataset demographic is such that the depressed participants mostly came from households with generally lower socioeconomic status and mothers with higher levels of depressive symptoms[24]. This reflects associations between adolescent depression with lower socioeconomic status and maternal depression [7].

Therefore, it would need to be validated with information relating to the parents diagnosis to build dependent models. An alternative could be to additionally analyze the parent’s speech independent of adolescent conversations.
Chapter Eight:

SUMMARY AND FUTURE WORK

8.1 Preview
This chapter, summarizes results, provides an overall conclusion to the thesis and discusses major findings including answers to research questions in Section 8.2. This thesis investigated two main research areas: emotion (recognition and prediction) and depression detection (acoustic and CMS approaches) discussed in Sections 8.2.1 and 8.2.2 respectively.

In addition this chapter provides an account of future work, in Section 8.3, that relates to the emotion and depression studies. The future work outlined includes a summary of further improvements and extensions of the work described in this thesis.

8.2 Discussion and Summary
Depression was the focal point of this research and was supplemented by emotion related studies ultimately aimed towards an integrated emotion and depression recognition system.

Emotion and emotion regulation are important concepts facilitating an effective human-machine communication and are psychologically linked to depression. Depression research has been motivated by the need for a new objective and fully automatic detection system capable of analyzing depression risk factors.
8.2.1 Emotion Related Studies

8.2.1.1 Automatic Emotion Recognition (AER) Studies

The emotion related studies have utilized an emotional speech database (EMO-DB) with seven emotions (anger, bored, disgust, fear, neutral, sad, joy). The differentiation between emotional classes was based on acoustic parameters extracted from both glottal (G-MFCC and G-TEO) and speech (S-MFCC and S-TEO) waveforms.

The initial binary and multiclass AER experiments used standard 1D features applied to NN, GMM and SVM classifiers. Further work investigated new 2D features applied to a DNN and the proposed OMC-DNN.

The study outcomes provided the following answers to the AER research questions previously outlined in Section 1.3:

Q1. What features are most effective for binary and multiclass AER?
A literature review showed that, MFCC and TEO parameters led to promising results in past AER studies. Experimentation determined that, MFCC were generally more effective than TEO parameters in both binary and multiclass AER.

The TEO parameters extracted from the glottal waveform (G-TEO) outperformed TEO parameters extracted from the speech waveform (S-TEO). However, the G-MFCC did not perform as well as the S-MFCC. These observations were consistent for all classifiers (i.e. SVM, NN, GMM) and genders (i.e. GIM, GDM-M and GDM-F).
Q2. How does a new deep learning approach compare to benchmark classifiers?

A comparison of standard 1D acoustic feature vectors with benchmark classifiers (NN, GMM and SVM) showed that, the DNN, using new 2D features performed significantly better.

The shallow NN struggled with the 2D representation of acoustic speech features, indicating that, the shallow NN may not have sufficient capacity to learn abstract concepts embedded into image arrays. Another explanation could be the due to the change in feature dimension size (i.e. 2D image stacked as a vector) regardless of the conceptualization of the 1D or 2D feature type.

Q3. Can an optimized multichannel DNN improve emotion recognition performance compared to the benchmark classification systems?

This thesis proposed a new AER approach implementing an optimized multichannel DNN (OMC-DNN) with 2D representation of speech features in weighted SC-DNN. It was shown that the OMC-DNN provided significant improvements of AER accuracy compared to individual SC-DNN for all binary classification tests and to a lesser extent multiclass AER.

The best overall AER results were obtained with the OMC-DNN with the following 2D features in a four-channel set: S-TEO, S-MFCC, G-TEO and G-MFCC. This configuration led to an average accuracy of 91% for the GIM, 93% for GDM-F and 96% for GDM-M.
8.2.1.2 Affect Prediction Studies

A range of methods was investigated towards efficient prediction of speakers’ emotional states from knowledge of their past states during dyadic conversations. Prediction was investigated using Traditional first order biased Random Walks (RW) and extended to higher order RW (HORW) based on a single speaker only.

To improve upon the RW and HORW, a new interacting RW (IRW) incorporating Influence Coefficients resulting from the HOEIM was proposed. This approach predicts speaker states using two speakers and their mutual influences.

Additionally, a 2D visualization of the RW was proposed to facilitate easier tracking of emotion state transitions and emotional flow during conversations. Tracking is based on the defined emotional drifts as a means to analyze a subject’s emotional transitions.

In this study the patterns were observed in general in relation to depression indicators. This could be extended to severity levels and to determine the efficacy of treatment with before and after comparisons of the 2D RW.

An alternative approach to emotion prediction was based on non-linear autoregressive models (NAR). The NAR included first-order and higher-order models implemented with ANFIS and NN and compared to the RW and HORW.

To incorporate effects from a conversational partner, first-order and higher-order NARX models, with NN and ANFIS, were validated and compared to IRW.
The study outcomes provided the following answers to the emotion prediction research questions previously outlined in Section 1.3:

**Q1. Is it possible to predict/forecast emotion of adolescents in conversations based on a memory of past states?**
Experiments described in the thesis have shown that, it is possible to predict positive, negative, neutral and silent constructs from past states of conversations using various RW and NAR modeling approaches.

**Q2. Do higher order memories improve the performance of affect prediction?**
Extending the RW by including multiple past states with higher order conditional probabilities improved affect prediction performance. Similarly, the NAR approaches were extended into higher order by introducing tapped time delay at the input and increased test classification accuracy.

**Q3. How to incorporate dependencies/influences between speakers towards improving affect prediction.**
It was shown that the intra- and inter-speaker dependencies can be incorporated using the influence coefficients (ICs) of the EIM. The ICs extracted from the EIM, weighted the conditional probabilities of adolescent and parent to create an IRW to predict the child’s next state and improved on the simple RW. The NAR-ANFIS and NAR-NN were extended to include memory of the parent’s past states as an exogenous input to create NARX-ANFIS and NARX-NN and increased prediction performance.
8.2.2 Depression Related Studies

The depression research used a subset of dyadic parent-adolescent conversations (29 depressed and 34 control) from the ORI-DB and is accompanied by second-by-second LIFE annotations. This information enables detection of depression using the audio signal or emotion annotations (first automated depression study from this viewpoint).

To determine if a subject is depressed or control two main approaches have been investigated: acoustic features (speech signal) and Influence Model features (LIFE annotations). In both cases extracted features were used to train machine-learning models to classify subjects or utterances as depressed or non-depressed.

8.2.2.1 Conversation Modeling System (CMS) Approach

The aim of the CMS approach was to develop a model to analyze emotional influences during adolescent-parent conversations with an application in detection depression. This was achieved with various EIM that generate features that qualitatively and quantitatively analyze risk factors and detect depression.

The study outcomes provided the following answers to the CMS depression detection research questions previously outlined in Section 1.3:

Q1. What is the most capable and efficient model of emotional interactions within conversations?

Based on psychological and social analysis studies it was determined that emotional influences are important in depressed patient’s family interactions. It was proposed to use dyadic conversation emotional construct annotations (LIFE) to generate parameters from an new emotional Influence Model (EIM).
Additionally, extensions of the EIM to include a trajectory of time delays as a Dynamic Influence Model and a multiple past states as a Higher Order Influence Model (HOEIM) was proposed.

Comparing the Log-Likelihood of various EIM implementations found DEIM generally performed worse as the delay increased. In contrast the HOEIM improved as the order increased compared to the original 1st order EIM. The MMM approach to higher order modeling was the worst HOEIM and KN-ngram approach was the best.

Q2. How to optimally fit the new conversation modeling system, Influence Model, into the data efficiently?

A preliminary examination was conducted comparing multiple optimization algorithms. It was determined the resultant parameters for all approaches, (SA, QN, MVR, GD, SA_GA, EM), gave statistically insignificant differences. The EM was significantly quicker and was given as the idyllic optimization algorithm.

Q3. Can the CMS provide psychologically valid qualitative interoperations and quantitative statistical significance?

The learned EIM parameters (transition probabilities and IC) provided quantitative parameters that showed statistical significances between depressed and non-depressed adolescents. Qualitative interoperations were consistent with current psychological explanations in relation to parent-adolescent relationships and depression.

The dynamic (DEIM) and higher order (HOEIM) emotional influence models led to more statistically significant features and arguably more detailed interpretations of time related emotion dependencies in conversations.
Q4. **How to most efficiently automatically detect depression from the CMS?**

The parameters learned from the original EIM \((t-I)\) attained on average around 70% accuracy with an SVM, 79% with NN, 87% with an optimized NN and 83% using a 2-stage mRMR. Further investigation was required to fully understand the intricacies of conversations beyond a delay step \((t-I)\).

The DEIM improved depression classification performance with a top \((t-5/FCl)\) of 97% with 99.1%/93.3% sensitivity specificity ratio with an O-NN. The HOEIM was the greatest EIM variant and best at higher orders. The HOEIM-MMM had lower classification compared to the HOEIM-KN-ngram approach.

The overall top result was attained using 2-stage optimized feature selection method and SVM classifier with the HOEIM-KN-ngram parameters \((4^{th} \text{ order/PSI})\) with 98.9% accuracy and 99%/98.7% sensitivity to specificity ratio.

8.2.2.2 Acoustic Speech Feature modeling Approach

The acoustic feature approach investigated six feature categories \((P, S, G, GW, C, T)\) for adolescent depression detection. Initial examination on individual feature sets, from subcategories, was extended to grouping categorical features and combinations. This proceeded into new optimized spectral roll-off features and glottal waveform features proposed for depression detection.

The study outcomes provided the following answers to the acoustic depression detection research questions previously outlined in Section 1.3:
Q1. Is it possible to detect adolescent depression from speech parameters during family conversations?

The experiments have confirmed that it is possible to efficiently detect depression in adolescents using speech parameters from dyadic parent-adolescent spontaneous conversations. A range of acoustic features derived from physiologically related characteristics, including Spectral (S), Prosodic (P), glottal (G), glottal waveform (GW), cepstral (C), TEO (T), detected depression.

Q2. What acoustic features show significant differences and provide discriminating risk factors related to depression in adolescents?

A total of six feature categories: S, P, G, GW, C and T were examined and with MANOVA and the subcategories analyzed with ANOVA. It was found that for males, females and all three topics all of the individual features were statistically significant ($p<0.05$) except jitter and shimmer from the prosodic (P) group.

Q3. What is the optimal combination of features and classification setup towards achieving the most efficient depression detection performance?

A preliminary investigation, with $F0$ as an acoustic feature, found that gender plays an important role in depression detection. Depression accuracy was higher using gender dependent models (GDM-M and GDM-F) compared to gender independent (GIM) and for most acoustic features the GDM-M was more accurate than GDM-F.

The new optimized sub-set from a proposed spectral roll-off range, using a 2-stage mRMR/SVM, improved depression classification compared to the entire feature set and individual coefficients used in past studies.
SVM depression classification was examined with each sub-category feature set and found that of the six categories (G, P, S, GW, T, C), the S category was on average the best with the inclusion of the new optimized spectral roll-offs. Without the optimized roll-off, replicating past studies (i.e. S* category) was considerably worse than the new S category. The next best category was the G category using both GTD and GFD feature sets and the worst category was the prosodic set (P).

MFCC and TEO parameters derived from the glottal waveform improved on the corresponding speech waveform features in the GDM-M. In the GDM-F case only the G-TEO was useful and not G-MFCC (i.e. MFCC>G-MFCC for GDM-F).

The highest average accuracy achieved, across three topics, using acoustic features was 98.8% using P+S for GDM-M with a top of 98.9% accuracy (99.9%/98.0% sensitivity/specificity) for PSI. In GDM-F case an average of 97.1% using G+S and a best of 97.9% accuracy (99.5%/96.2% sensitivity/specificity) in EPI.

Q4. Can depression in adolescents be detected by analyzing the acoustic speech features from their parent’s speech during family interactions?

One of the most interesting findings of this thesis was the observation of significant differences between acoustic speech characteristics of parents of depressed and non-depressed adolescents. Therefore it was possible that parent’s acoustic features could provide valuable information in regards to depression of their children. SVM classification, with six categorical feature sets, in GDM-M had between 53.6% (FCI/MFCC) and 71.2% (PSI/S) and GDM-F between 52.5% (FCI/MFCC) and 73.1% (PSI/S).
8.2.3 Acoustic Features versus Conversation Modeling (EIM) Approaches

This thesis introduced a new approach to depression detection using features extracted from a CMS implemented with an Emotional Influence Model (EIM) and proposed extensions (DEIM and HOEIM), based on the LIFE annotations of the ORI-DB.

This was compared to a traditional acoustic feature extraction approach using speech from the same ORI-DB. A range of features and categories (i.e. S, P, G, C, T) commonly used in depression literature were investigated. New acoustic features based on the spectral roll-off range were optimized with feature selection. An examination of previously used G-TEO and new G-MFCC were compared to speech waveform features.

Table 8-1 provides a ranked list of all feature/classification methods tested in CMS and acoustic depression detection approaches. Dark grey indicates the CMS approach with new EIM (HOEIM and DEIM) parameters, light grey denotes features introduced from an acoustic speech feature approach and the cases without shading are acoustic features previously used in depression literature.
Table 8-1 Ranking of the Depression Classification Performance of the Average Accuracy (%) of Each Topic (EPI, FCI, PSI) and Gender Model (GDM-M and GDM-F) using the ORI-DB Speech Signals for the Acoustic Feature Approach and the LIFE Construct Annotations (positive, negative, neutral, silence) for the Influence Model Approach. The dark shaded cells denote the newly proposed EIM features, lighter shaded cells indicate the speech acoustic features that include new features and clear cells are speech acoustics used previously.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Model Feature</th>
<th>Order/Delay</th>
<th>Classifier</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>P+S</td>
<td></td>
<td>SVM</td>
<td>97.8</td>
</tr>
<tr>
<td>2</td>
<td>G-TEO+S</td>
<td></td>
<td>SVM</td>
<td>96.1</td>
</tr>
<tr>
<td>3</td>
<td>G+S</td>
<td></td>
<td>SVM</td>
<td>95.7</td>
</tr>
<tr>
<td>4</td>
<td>G-MFCC+S</td>
<td></td>
<td>SVM</td>
<td>95.6</td>
</tr>
<tr>
<td>5</td>
<td>HOEIM-KN-ngram</td>
<td>4th</td>
<td>mRMR- SVM</td>
<td>95.5</td>
</tr>
<tr>
<td>6</td>
<td>Spectral (S)</td>
<td></td>
<td>SVM</td>
<td>94.9</td>
</tr>
<tr>
<td>7</td>
<td>TEO+S</td>
<td></td>
<td>SVM</td>
<td>94.5</td>
</tr>
<tr>
<td>8</td>
<td>DEIM</td>
<td>t-5</td>
<td>O-NN</td>
<td>93.4</td>
</tr>
<tr>
<td>9</td>
<td>DEIM</td>
<td>t-6</td>
<td>O-NN</td>
<td>92.3</td>
</tr>
<tr>
<td>10</td>
<td>MFCC+S</td>
<td></td>
<td>SVM</td>
<td>92.1</td>
</tr>
<tr>
<td>11</td>
<td>DEIM</td>
<td>t-4</td>
<td>NN</td>
<td>90.8</td>
</tr>
<tr>
<td>12</td>
<td>DEIM</td>
<td>t-4</td>
<td>O-NN</td>
<td>90.8</td>
</tr>
<tr>
<td>13</td>
<td>DEIM</td>
<td>t-2</td>
<td>O-NN</td>
<td>89.6</td>
</tr>
<tr>
<td>14</td>
<td>G-TEO+G</td>
<td></td>
<td>SVM</td>
<td>89.5</td>
</tr>
<tr>
<td>15</td>
<td>DEIM</td>
<td>t-3</td>
<td>O-NN</td>
<td>88.8</td>
</tr>
<tr>
<td>16</td>
<td>TEO+G</td>
<td></td>
<td>SVM</td>
<td>88.5</td>
</tr>
<tr>
<td>17</td>
<td>DEIM</td>
<td>t-1</td>
<td>O-NN</td>
<td>86.6</td>
</tr>
<tr>
<td>18</td>
<td>P+G</td>
<td></td>
<td>SVM</td>
<td>86.0</td>
</tr>
<tr>
<td>19</td>
<td>HOEIM-KN-ngram</td>
<td>3rd</td>
<td>mRMR- SVM</td>
<td>85.5</td>
</tr>
<tr>
<td>20</td>
<td>HOEIM-KN-ngram</td>
<td>5th</td>
<td>mRMR- SVM</td>
<td>85.2</td>
</tr>
<tr>
<td>21</td>
<td>DEIM</td>
<td>t-8</td>
<td>O-NN</td>
<td>85.0</td>
</tr>
<tr>
<td>22</td>
<td>DEIM</td>
<td>t-5</td>
<td>NN</td>
<td>84.3</td>
</tr>
<tr>
<td>23</td>
<td>DEIM</td>
<td>t-3</td>
<td>NN</td>
<td>83.6</td>
</tr>
<tr>
<td>24</td>
<td>DEIM</td>
<td>t-10</td>
<td>O-NN</td>
<td>83.5</td>
</tr>
<tr>
<td>25</td>
<td>HOEIM-KN-ngram</td>
<td>2nd</td>
<td>SVM</td>
<td>82.9</td>
</tr>
<tr>
<td>26</td>
<td>HOEIM-KN-ngram</td>
<td>1st</td>
<td>mRMR- SVM</td>
<td>82.9</td>
</tr>
<tr>
<td>27</td>
<td>GT+GF</td>
<td></td>
<td>SVM</td>
<td>82.2</td>
</tr>
<tr>
<td>28</td>
<td>DEIM</td>
<td>t-8</td>
<td>NN</td>
<td>82.2</td>
</tr>
<tr>
<td>29</td>
<td>G-MFCC+G</td>
<td></td>
<td>SVM</td>
<td>81.9</td>
</tr>
<tr>
<td>30</td>
<td>DEIM</td>
<td>t-6</td>
<td>NN</td>
<td>81.8</td>
</tr>
<tr>
<td>31</td>
<td>DEIM</td>
<td>t-9</td>
<td>O-NN</td>
<td>80.5</td>
</tr>
<tr>
<td>32</td>
<td>MFCC+G</td>
<td></td>
<td>SVM</td>
<td>79.7</td>
</tr>
<tr>
<td>33</td>
<td>DEIM</td>
<td>t-7</td>
<td>O-NN</td>
<td>79.4</td>
</tr>
<tr>
<td>34</td>
<td>DEIM</td>
<td>t-1</td>
<td>NN</td>
<td>79.3</td>
</tr>
<tr>
<td>35</td>
<td>DEIM</td>
<td>t-10</td>
<td>NN</td>
<td>78.8</td>
</tr>
<tr>
<td>36</td>
<td>DEIM</td>
<td>t-2</td>
<td>NN</td>
<td>78.3</td>
</tr>
<tr>
<td>37</td>
<td>HOEIM-KN-ngram</td>
<td>3rd</td>
<td>SVM</td>
<td>76.8</td>
</tr>
<tr>
<td>38</td>
<td>HOEIM-KN-ngram</td>
<td>5th</td>
<td>SVM</td>
<td>76.5</td>
</tr>
<tr>
<td></td>
<td>Feature</td>
<td>Stage</td>
<td>Classifer</td>
<td>Accuracy</td>
</tr>
<tr>
<td>---</td>
<td>-----------------</td>
<td>-------</td>
<td>-----------</td>
<td>----------</td>
</tr>
<tr>
<td>39</td>
<td>Roll-off</td>
<td>-</td>
<td>SVM</td>
<td>76.4</td>
</tr>
<tr>
<td>40</td>
<td>HOEIM-KN-ngram</td>
<td>4th</td>
<td>SVM</td>
<td>75.9</td>
</tr>
<tr>
<td>41</td>
<td>G-TEO</td>
<td>-</td>
<td>SVM</td>
<td>75.3</td>
</tr>
<tr>
<td>42</td>
<td>GTD</td>
<td>-</td>
<td>SVM</td>
<td>75.1</td>
</tr>
<tr>
<td>43</td>
<td>HOEIM-MMM</td>
<td>3rd</td>
<td>SVM</td>
<td>75.0</td>
</tr>
<tr>
<td>44</td>
<td>HOEIM-KN-ngram</td>
<td>2nd</td>
<td>SVM</td>
<td>74.9</td>
</tr>
<tr>
<td>45</td>
<td>Prosodic (P)</td>
<td>-</td>
<td>SVM</td>
<td>73.2</td>
</tr>
<tr>
<td>46</td>
<td>DEIM</td>
<td>t-7</td>
<td>NN</td>
<td>72.8</td>
</tr>
<tr>
<td>47</td>
<td>HOEIM-MMM</td>
<td>2nd</td>
<td>SVM</td>
<td>72.7</td>
</tr>
<tr>
<td>48</td>
<td>Spectral* (S*)</td>
<td>-</td>
<td>SVM</td>
<td>71.0</td>
</tr>
<tr>
<td>49</td>
<td>HOEIM-MMM</td>
<td>4th</td>
<td>SVM</td>
<td>70.6</td>
</tr>
<tr>
<td>50</td>
<td>TEO</td>
<td>-</td>
<td>SVM</td>
<td>70.4</td>
</tr>
<tr>
<td>51</td>
<td>HOEIM-KN-ngram</td>
<td>1st</td>
<td>SVM</td>
<td>70.0</td>
</tr>
<tr>
<td>52</td>
<td>G-MFCC</td>
<td>-</td>
<td>SVM</td>
<td>70.0</td>
</tr>
<tr>
<td>53</td>
<td>PSD</td>
<td>-</td>
<td>SVM</td>
<td>69.9</td>
</tr>
<tr>
<td>54</td>
<td>DEIM</td>
<td>t-9</td>
<td>NN</td>
<td>69.7</td>
</tr>
<tr>
<td>55</td>
<td>MFCC</td>
<td>-</td>
<td>SVM</td>
<td>69.6</td>
</tr>
<tr>
<td>56</td>
<td>HOEIM-MMM</td>
<td>5th</td>
<td>SVM</td>
<td>67.6</td>
</tr>
<tr>
<td>57</td>
<td>Fo</td>
<td>-</td>
<td>SVM</td>
<td>67.5</td>
</tr>
<tr>
<td>58</td>
<td>Formants</td>
<td>-</td>
<td>SVM</td>
<td>67.5</td>
</tr>
<tr>
<td>59</td>
<td>DEIM</td>
<td>t-1</td>
<td>SVM</td>
<td>67.1</td>
</tr>
<tr>
<td>60</td>
<td>HOEIM-MMM</td>
<td>1st</td>
<td>SVM</td>
<td>67.1</td>
</tr>
<tr>
<td>61</td>
<td>DEIM</td>
<td>t-5</td>
<td>SVM</td>
<td>67.1</td>
</tr>
<tr>
<td>62</td>
<td>DEIM</td>
<td>t-2</td>
<td>SVM</td>
<td>67.1</td>
</tr>
<tr>
<td>63</td>
<td>DEIM</td>
<td>t-3</td>
<td>SVM</td>
<td>66.7</td>
</tr>
<tr>
<td>64</td>
<td>Spectral (S) Parent</td>
<td>-</td>
<td>SVM</td>
<td>66.4</td>
</tr>
<tr>
<td>65</td>
<td>GFD</td>
<td>-</td>
<td>SVM</td>
<td>66.4</td>
</tr>
<tr>
<td>66</td>
<td>DEIM</td>
<td>t-4</td>
<td>SVM</td>
<td>65.7</td>
</tr>
<tr>
<td>67</td>
<td>DEIM</td>
<td>t-7</td>
<td>SVM</td>
<td>65.6</td>
</tr>
<tr>
<td>68</td>
<td>DEIM</td>
<td>t-9</td>
<td>SVM</td>
<td>65.4</td>
</tr>
<tr>
<td>69</td>
<td>DEIM</td>
<td>t-6</td>
<td>SVM</td>
<td>65.3</td>
</tr>
<tr>
<td>70</td>
<td>DEIM</td>
<td>t-8</td>
<td>SVM</td>
<td>65.1</td>
</tr>
<tr>
<td>71</td>
<td>Prosodic (P) Parent</td>
<td>-</td>
<td>SVM</td>
<td>64.6</td>
</tr>
<tr>
<td>72</td>
<td>G-TEO (GW) Parent</td>
<td>-</td>
<td>SVM</td>
<td>64.5</td>
</tr>
<tr>
<td>73</td>
<td>DEIM</td>
<td>t-10</td>
<td>SVM</td>
<td>62.7</td>
</tr>
<tr>
<td>74</td>
<td>GT+GF (G) Parent</td>
<td>-</td>
<td>SVM</td>
<td>62.1</td>
</tr>
<tr>
<td>75</td>
<td>LogE</td>
<td>-</td>
<td>SVM</td>
<td>60.4</td>
</tr>
<tr>
<td>76</td>
<td>TEO (T) Parent</td>
<td>-</td>
<td>SVM</td>
<td>59.4</td>
</tr>
<tr>
<td>77</td>
<td>G-MFCC (GW) Parent</td>
<td>-</td>
<td>SVM</td>
<td>58.8</td>
</tr>
<tr>
<td>78</td>
<td>Jitter</td>
<td>-</td>
<td>SVM</td>
<td>58.7</td>
</tr>
<tr>
<td>79</td>
<td>Entropy</td>
<td>-</td>
<td>SVM</td>
<td>56.0</td>
</tr>
<tr>
<td>80</td>
<td>Centroid</td>
<td>-</td>
<td>SVM</td>
<td>54.2</td>
</tr>
<tr>
<td>81</td>
<td>MFCC (C) Parent</td>
<td>-</td>
<td>SVM</td>
<td>54.2</td>
</tr>
<tr>
<td>82</td>
<td>Flux</td>
<td>-</td>
<td>SVM</td>
<td>53.8</td>
</tr>
<tr>
<td>83</td>
<td>Shimmer</td>
<td>-</td>
<td>SVM</td>
<td>53.0</td>
</tr>
</tbody>
</table>
The top features are mostly acoustic but have downsides, such as reliance on audio quality, gender dependence and only provide quantitative analysis. In contrast, the newly proposed conversation modeling techniques provide the unique quantitative parameters with corresponding qualitative descriptions of depression and emotions.

All of the acoustic features that outperform the proposed EIM approach include the new optimized spectral roll-off range. The EIM and variations outperform the standard acoustic features. The top 15% of the 83 different feature/classification cases are new features proposed in this thesis and outperform classical features.

It should also be noted that the experimental results attained using the classifiers with optimized parameters (e.g. nodes in NN) or features (using mRMR) may be optimistic as the parameters were tuned during preliminary experiments with some knowledge of the training data.

8.2.4 Performance overview of depression classification compared to past literature

Direct comparison of the depression and emotion related studies to past literature is difficult with several problems to consider as follows:

— A fundamental problem is inconsistent datasets used between studies such as different recording conditions. Additionally, the style of speech (spontaneous, simulated, acted) can alter levels of arousal and the ease of differentiation.

— Inconsistencies between definition of emotions, feature/classifier setups and subject and gender independent/dependent models are difficult to compare.

— Dissimilar performance measures (accuracy, recall, F-score, error rates) and different evaluation setups (subject-based, utterance, words, sentence, frame).
Different depression modeling approaches, features type, performance measures and datasets, makes directly compare our results to other studies difficult. It is still important to compare the new EIM features and acoustic features with literature. Table 8-2 provides a selective summary of the most similar, recent and effective past depression detection studies ranked on the best feature accuracy attained in the respective study.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>France, et al. (2000) [109]</td>
<td>PSD</td>
</tr>
<tr>
<td>96% GDM-F</td>
<td></td>
</tr>
<tr>
<td>Alghowinem, et al. (2013) [421]</td>
<td>P+S+MFCC</td>
</tr>
<tr>
<td>Ozdas, et al. (2009) [106]</td>
<td>Glottal slope+jitter</td>
</tr>
<tr>
<td>*Low, et al. (2011) [140]</td>
<td>TEO</td>
</tr>
<tr>
<td>79% GDM-F</td>
<td></td>
</tr>
<tr>
<td>*Low, et al. (2010) [108]</td>
<td>TEO+P+S</td>
</tr>
<tr>
<td>75% GDM-F</td>
<td></td>
</tr>
<tr>
<td>65% GDM-F</td>
<td></td>
</tr>
<tr>
<td>Ooi, et al. (2012) [65]</td>
<td>Glottal</td>
</tr>
<tr>
<td>60% GDM-F</td>
<td></td>
</tr>
</tbody>
</table>

Comparing the absolute optimal results achieved in each study with the top in this thesis from the acoustic and conversation modeling approaches as follows:

- Acoustic features: P+S (GDM-M) 98.8% and G+S (GDM-F) 97.1%
- Conversation Modeling: HOEIM-KN-ngram mRMR/SVM (GIM) 95.5%

The acoustic and CMS approaches in this thesis performed equal or better, if only slightly, compared to the past literature. The best results attained in past studies in the GIM case was 94% by France et al. [109] and in the GDM case 96% (GDM-F) and 91% (GDM-M) by Moore et al. [105].


8.3 Future Work

This research has shown it is possible to detect adolescent clinical depression using acoustic speech and with a new CMS approach. However, automatic depression detection is still a difficult task with many factors that could potentially contribute to depression development (genetic, environment, social and psychological) [24].

There are various limitations in this research related to the database and could benefit from validation with other databases and for cross corpus capabilities. Furthermore work can be conducted following on from the results and methods presented in this thesis and could consider the following ideas:

1) A larger audio-visual database especially for certain gender pairing (i.e. the small father-son combination). Furthermore a wider range of ages would be necessary to extend this into adults depression detection, due to the notable difference in adult and child speech and depressive symptoms [498].

2) Continue on from the preliminary adolescent depression detection examination from parent’s speech with a more ideal database including the parents’ depressive status. Therefore, multiple parent depression-status dependent models could be built to detect adolescent depression.

3) The IM could be applied to another annotation system used independently or in conjunction with multiple annotation sequences (i.e. coupled chains). Intuitively the EIM would be extended past dyadic conversations.

4) Extend to possibly include facial images or biological signals (e.g. EEG, ECG, skin impedance) along with the acoustic speech analysis and EIM presented here in a multichannel system should be considered by future studies.
5) The scope of this thesis is limited to depression detection and future work could consider different mental health disorders or conversation analysis applications.

6) It would be good extend the use of the depression detection methods to determine the severity of depression using a standard depression scale rating.


HINTON, G., “Products of Experts”, in Proc. Of the 9th Int. Conf. on Artificial Neural Networks (ICANN), 1999, pp. 1-6.


TAYLOR, G., AND HINTON, G., “Factored conditional restricted Boltzmann machines for modeling motion style,” in Proc. of the 26th Int. Conf. on Machine Learning (ICML), Montreal, Canada, 2009.


299


Kitchener, B., and Jorm, A., Youth Mental Health First Aid: A manual for adults assisting youth. ORYGEN Research Centre, Melbourne, 2009.


Kraepelin, E., Manic-depressive Insanity and Paranoia, Edinburgh: Livingstone, 1921


306
Proceedings of the 22nd European Signal Processing Conference (EUSIPCO), 2014.


BONE., K “Acoustic-Prosodic and Turn-Taking Features in Interactions with Children with Neurodevelopmental Disorders”, in Proc. of Interspeech, 2016.


ÉViben, F., and SCHULLER, B., "openSMILE: The Munich Open-Source Large-scale Multimedia Feature Extractor", ACM SIGMM Records, vol.6, no.4, pp. 4-12, 2014

