Automated Screening Methods for Mental and Neuro-developmental Disorders

A thesis submitted in partial fulfilment
of the requirement for the degree of Doctor of Philosophy
Computer Science & Informatics

Syed, Mohammed Zafi Sherhan Shah

September 2018

Cardiff University
School of Computer Science & Informatics

Abstract

Mental and neuro-developmental disorders such as depression, bipolar disorder, and autism spectrum disorder (ASD) are critical healthcare issues which affect a large number of people. Depression, according to the World Health Organisation, is the largest cause of disability worldwide and affects more than 300 million people. Bipolar disorder affects more than 60 million individuals worldwide. ASD, meanwhile, affects more than 1 in 100 people in the UK. Not only do these disorders adversely affect the quality of life of affected individuals, they also have a significant economic impact.

While brute-force approaches are potentially useful for learning new features which could be representative of these disorders, such approaches may not be best suited for developing robust screening methods. This is due to a myriad of confounding factors, such as the age, gender, cultural background, and socio-economic status, which can affect social signals of individuals in a similar way as the symptoms of these disorders. Brute-force approaches may learn to exploit effects of these confounding factors on social signals in place of effects due to mental and neuro-developmental disorders.

The main objective of this thesis is to develop, investigate, and propose computational methods to screen for mental and neuro-developmental disorders in accordance with descriptions given in the Diagnostic and Statistical Manual (DSM). The DSM manual is a guidebook published by the American Psychiatric Association which offers common language on mental disorders. Our motivation is to alleviate, to an extent, the possibility of machine learning algorithms picking up one of the confounding factors to optimise performance for the dataset — something which we do not find uncommon in research literature.

To this end, we introduce three new methods for automated screening for depression from audio/visual recordings, namely: turbulence features, craniofacial movement features, and Fisher Vector based representation of speech spectra. We surmise that psychomotor changes due to depression lead to uniqueness in an individual's speech pattern which manifest as sudden and erratic changes in speech feature contours. The efficacy of these features is demonstrated as part of our solution to Audio/Visual Emotion Challenge 2017 (AVEC 2017) on Depression severity prediction. We also detail a methodology to quantify specific craniofacial movements, which we hypothesised could be indicative of psychomotor retardation, and hence depression. The efficacy of

craniofacial movement features is demonstrated using datasets from the 2014 and 2017 editions of AVEC Depression severity prediction challenges. Finally, using the dataset provided as part of AVEC 2016 Depression classification challenge, we demonstrate that differences between speech of individuals with and without depression can be quantified effectively using the Fisher Vector representation of speech spectra.

For our work on automated screening of bipolar disorder, we propose methods to classify individuals with bipolar disorder into states of remission, hypo-mania, and mania. Here, we surmise that like depression, individuals with different levels of mania have certain uniqueness to their social signals. Based on this understanding, we propose the use of turbulence features for audio/visual social signals (i.e. speech and facial expressions). We also propose the use of Fisher Vectors to create a unified representation of speech in terms of prosody, voice quality, and speech spectra. These methods have been proposed as part of our solution to the AVEC 2018 Bipolar disorder challenge.

In addition, we find that the task of automated screening for ASD is much more complicated. Here, confounding factors can easily overwhelm socials signals which are affected by ASD. We discuss, in the light of research literature and our experimental analysis, that significant collaborative work is required between computer scientists and clinicians to discern social signals which are robust to common confounding factors.

Acknowledgements

I express my deepest gratitude and appreciation to my supervisors from Cardiff School of Computer Science and Informatics, Prof. David Marshall and Dr. Kirill Sidorov, both of whom are exceptionally nice human beings. They have supported me throughout this endeavour with enlightening discussions and made administrative processes simpler to follow. I especially appreciate their belief and their trust as I undertook research on the development of automated screening methods for mental and neuro-developmental disorders.

I am thankful to Dr. Catherine Jones from Cardiff School of Psychology for all her help and support, including granting us access to the dataset collected by her group. This enabled us to work on automated screening of autism spectrum disorder based on speech modality alone.

I would also like to thank my good friend and fellow PhD student, Aled Owen. He has been like a brother to me in a country thousands of miles away from my own. I am thankful for all the support he has offered me during my PhD in Cardiff. I am also grateful to have had the opportunity to discuss my research with him and receive critical feedback.

I would like to thank Helen Williams, Khtam Al Meyah, Kaelon Lloyd, Tom Hartley, Ioannis Kaloskampis, Julien Schroeter, and Joseph Redfern for all their help and support.

Finally, I would like to thank the most important people in this world for me, which is my family. My father Sher Mohammed, my mother Naheed, my wife Hina, my brothers Shehram and Abbas, my sister Fiza, my sister-in-law Madiha, and my cute little niece Maryam Fatima. I am grateful for their support all the way from Hyderabad, Sindh, Pakistan.

Dedication

– To the memory of my late grandmother, Shireen Shah. I was one of her carers after she suffered from dementia and Alzheimer's disease. It was that experience which motivated me to pursue research towards the development of computerised methods to understand human behaviour, especially of those individuals who suffer from mental and neuro-developmental disorders or cognitive impairments.

– To my parents, Sher Mohammed Shah and Naheed Shah. They have always been an inspiration for me by demonstrating the value of hard and honest work.

Contents

List of Publications

Published

- Zafi Sherhan Syed, Kirill Sidorov, and David Marshall, "Automated Screening for Bipolar Disorder from Audio/Visual Modalities", ACM International Workshop on Audio/Visual Emotion Challenge (AVEC), pp. 39-45, 2018.
- Zafi Sherhan Syed, Julien Schroeter, Kirill Sidorov, and David Marshall, "Computational Paralinguistics: Automatic Assessment of Emotions, Mood, and Behavioural State from Acoustics of Speech", The 19th Annual Conference of the International Speech Communication Association INTERSPEECH, pp. 511-515, 2018.
- Zafi Sherhan Syed, Kirill Sidorov, and David Marshall, "Depression Severity Prediction Based on Biomarkers of Psychomotor Retardation", ACM International Workshop on Audio/Visual Emotion Challenge (AVEC), pp. 37-43, 2017.
- Zafi Sherhan Shah, Kirill Sidorov, and David Marshall, "Psychomotor Cues for Depression Screening", IEEE International Conference on Digital Signal Processing (DSP), pp. 1-5, 2017.

In Preparation

- Zafi Sherhan Syed, Kirill Sidorov, Catherine Jones, and David Marshall, "Automated Screening for Autism Spectrum Disorder using Speech Modality", 2019.

List of Figures

List of Tables

List of Abbreviations

AAM	Active Appearance Model
ADHD	Attention Disorder Hyper Activity Disorder
ANEW	Affective Norms for English words
APA	American Psychiatric Association
ASD	Autism Spectrum Disorder
ATD	Atypically Developing
AU	Action Unit
AVEC	Audio/Visual Emotion Recognition Challenge
AVEC 2014 DSC	AVEC 2014 Depression severity prediction subchallenge
AVEC 2014 DSC–FF	AVEC 2014 DSC – Freeform task
AVEC 2014 DSC–NW	AVEC 2014 DSC – Northwind task
AVEC 2016 DCC	AVEC 2016 Depression classification subchallenge
AVEC 2017 DSC	AVEC 2017 Depression severity prediction subchallenge
AVEC 2018 BDS	AVEC 2018 Bipolar disorder subchallenge
BDI	Beck Depression Inventory
BoAW	Bag-of-Audio-Words
BoVW	Bag-of-Visual-Words
BoW	Bag-of-Words
BPRS	Brief Psychiatric Rating Scale
CCA	Canonical Correlation Analysis

CDC	Centers for Disease Control and Prevention
CERT	Computer Expression Recognition Toolbox
CNN	Convolutional Neural Networks
ComParE	COMutational PARalinguistics challengE
ComParE–AASC	ComParE 2018 Atypical-affect subchallenge
ComParE–CSC	ComParE 2018 Crying subchallenge
ComParE–SAAC	ComParE 2018 Self-assesed affect subchallenge
COVAREP	Cooperative VoiceAnalysis Repository for Speech Technologies
CPP	Cepstral Peak Prominance
CPSD	Child Pathological Speech Database
CQT	Constant-Q transform
DAIC	Distress Assessment Interview Corpus
DFT	Discrete Fourier Transform
DNN	Deep Neural Networks
DSM-5	Diagnostic and Statistical Manual of Mental Disorders
DSP	Digital Signal Processing
e2e	End to end
eGeMAPS	Extended Geneva Minimalistic Acoustic Parameter Set
ELM	Extreme Learning Machines
EmotAsS	Emotional Sensitivity Assistance System for People with Disabilities
F0	Pitch/fundamental frequency
FACS	Facial Action Coding System
FAU	Facial Action Unit
FFT	Fast Fourier Transform

FV	Fisher Vectors
GeMAPS	Geneva Minimalistic Acoustic Parameter Set
GEWELMs	Greedy Ensemble of Extreme Learning Machines
GLM	Generalised Linear Models
GMM	Gaussian Mixture Models
GRU	Gated Recurrent Units
GSR	Gaussian Staircase Regression
HAM-D	Hamilton Rating Scale for Depression
HNR	Harmonics to Noise Ratio
HOF	Histogram of Optical Flow
HOG	Histogram of Oriented Gradients
HRSD	Hamilton Rating Scale for Depression
IS10-Paralinguistics	Interspeech 2010 Paralinguistics
LBP	Local Binary Patterns
LBP-TOP	Local Binary Patterns over Three Orthogonal Planes
LIWC	Linguistic Inquiry and Word Count
LLDs	Low Level Descriptors
LOOCV	Leave One Out Cross Validation
LOSOCV	Leave One Subject Out Cross Validation
LPC	Linear Prediction Coefficients
LPQ	Local Phase Quantisation
LSTM	Long Short Term Memory
LTAS	Long Term Spectral Average
MAE	Mean Absolute Error
MCQ	Multiple Choice Questions
MDD	Major Depressive Disorder

MDQ	Maxima Dispersion Quotient
MFCCs	Mel Frequency Cepstral Coefficients
MHH	Motion History Histogram
MHI	Motion History Image
mRMR	Minimum Redundancy Maximum Relevance
MSE	Mean Square Error
NAQ	Normalized Amplitude Quotient
PCA	Principal Component Analysis
PCC	Pearson Correlation Coefficient
PHQ-8	Patient Health Questionnaire (8 item)
PHQ-9	Patient Health Questionnaire (9 item)
PLP	Perceptual Linear Prediction
PLS	Partial Least Squares
PSP	Parabolic Spectral Parameter
PTSD	Post Traumatic Stress Disorder
QIDS	Quick Inventory of Depressive Symptomatology
QIDS–C	Quick Inventory of Depressive Symptomatology (Clinician administered)
QIDS–SR	Quick Inventory of Depressive Symptomatology (Self Reported)
QOQ	Quasi-Open Quotient
RASTA	Relative Spectral Amplitude
RMSE	Root Mean Square Error
RVM	Relevance Vector Machine
SALAT	Suite of Automatic Linguistic Analysis Tools
SBS	Sequential Backward Search
SFS	Sequential Forward Search

SHS	Sub-Harmonic Summation
SRH	Summation of the Residual Harmonics
SSP	Social Signal Processing
SSP/AC	Social Signal Processing/Affective Computing
STFT	Short-Time Fourier transform
SVM	Support Vector Machine
SVR	Support Vector Regressor/Regression
TD	Typically Developing
UAR	Unweighted Average Recall
UBM	Universal Background Model
URTIC	Upper Respiratory Tract Infection Corpus
U-test	Mann-Whitney U-test
VAD	Voice Activity Detector
VLAD	Vectors of Locally Aggregated Descriptors
WEKA	Waikato Environment for Knowledge Analysis
WHO	World Health Organisation
WURSS	Wisconsin Upper Respiratory Symptom Survey
YMRS	Young Mania Rating Scale
ZCR	Zero Crossing Rate

Chapter 1

Introduction

In this chapter, we provide a gentle introduction to the field of Social Signal Processing (SSP) and briefly discuss its generic process flow diagram which can be followed to develop automated screening methods. We then summarise some of the outstanding challenges in the field and our motivation for pursuing research in social signal processing. We follow this by listing the contributions of our work and finish the chapter by presenting the outline for the rest of this thesis.

1.1 Social Signal Processing and Automated Screening

SSP is a relatively new research and technological domain which deals with automated extraction and processing of social signals with the aim to provide social intelligence to computers [3]. It is unique in the sense that it overlaps multiple disciplines such as computer science, engineering, and human sciences. SSP finds applications in augmented reality, gaming, marketing, and healthcare amongst many more, however, our focus in this thesis will be on the application of SSP in healthcare.

As the name suggests, SSP focuses on social signals exhibited by human beings. Let us therefore formally introduce social signals. Vinciarelli et al., who are credited with coining the term *social signal processing*, define a social signal as 'a communicative or informative signal that, either directly or indirectly, provides information about social facts, namely social interactions, social emotions, social attitudes, or social relations' [3,4]. A social interaction is based on exchange of communicative and informative signals. A communicative signal is the one in which a person tries to convey to another, where as an informative signal is one which is actually conveyed. According to Vinciarelli et al., social interactions are most meaningful when communicative and informative signals convey identical information. As we shall discuss later in this thesis, individuals

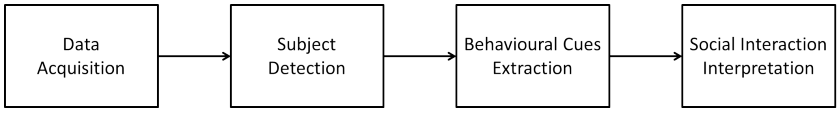


Figure 1.1: Generic process flow in social signal processing

who suffer from mental and neuro-developmental disorders often lack the ability to have meaningful social interactions since their social signals are impaired.

In its most generic form, SSP consists of the following set of procedures [5]: data acquisition, person detection, behavioural cues extraction, and social interaction interpretation, as illustrated in Figure 1.1. Here, the first two blocks are mostly concerned with computer science and engineering, whereas the last two require integration of knowledge from human sciences as well.

Data acquisition is the process of acquiring social signals which represent the social interaction between subject(s) and can be gathered using cameras, microphones, and wearable devices. Person detection is the process of discriminating between the intended subjects from the clutter within the acquired data. This requires the use of algorithms for facial recognition, speaker identification, tracking, and noise cancellation amongst others.

While the step of *person detection* enables one to collect data from the intended subject, one is often required to extract/compute relevant social signals for the task at hand. Here, background knowledge is often required from the field of human science so that only relevant social signals are extracted and stored for further processing. Some examples of most commonly used social signals include facial expressions, speech prosody, head movement, body posture, and hand gestures. Using these, many aspects of human behaviour can be inferred [4–6]. In a computational framework, one typically applies signal processing algorithms on raw data to extract these social signals. The final step is to train machine learning models for linking the extracted behavioural cues with meaningful interpretations from human sciences. For example, consider the task of emotion recognition where one can use background knowledge from Ekman's work on basic human emotions [7] to link facial expressions with emotions of anger, happiness, sadness, or surprise [8].

When SSP methods are used to identify behavioural symptoms of health related issues, the application of SSP falls into the category of *automated screening methods*. Unsurprisingly, SSP has quickly gained interest and recognition amongst researchers working in the fields of computer science, engineering, and human sciences especially for development of automated screening methods. In fact, from 2013 through 2017, annual competitions were organised for automated screening of depression using the platform of European Social Signal Processing Network (SSPNet) [9–14].

While the term SSP was popularised through SSPNet, especially by means of annual competitions under the umbrella of SSPNet, we note that this field

also exists under the name of Affective Computing (AC) [15], which predates SSP. In fact, we find that multiple researchers have attempted to propose different names for essentially the same field. For example, Narayanan et al. proposed the name 'Behavioral Signal Processing' in [16], whereas Valstar et al. [17] propose the name 'Behaviomedics'. In order to avoid further confusion between SSP and AC, in this thesis we shall refer to them jointly as SSP/AC.

1.2 Mental and Neuro-developmental Disorders

The Diagnostic and Statistical Manual of Mental Disorders (DSM) [18] is a reference book published by the American Psychiatric Association which offers a common language for the codification of mental disorders. The manual provides standard diagnostic criteria for various types of disorders, which include major depressive disorder (commonly known as depression), autism spectrum disorder (commonly known as autism), bipolar disorder, post-traumatic stress disorder, and schizophrenia amongst many others.

The DSM-5 manual (the latest edition) defines a mental disorder as a *'syndrome characterized by clinically significant disturbance in an individual's cognition, emotion regulation, or behaviour that reflects a dysfunction in the psychological, biological, or developmental processes underlying mental functioning'*. Meanwhile, *neuro-developmental disorders* are defined as an umbrella term for a group of conditions which start becoming apparent in the developmental phase of a child. According to the DSM-5, the neuro-developmental disorders are *'a group of conditions with onset in the developmental period. The disorders typically manifest early in development, often before the child enters grade school, and are characterized by developmental deficits that produce impairments of personal, social, academic, or occupational functioning. The range of developmental deficits varies from very specific limitations of learning or control of executive functions to global impairments of social skills or intelligence'*. Relevant to this thesis, depression and bipolar disorder are considered mental disorders, whereas ASD is considered a neuro-developmental disorder.

1.3 Problem Description

Conventional methods for screening of mental and neuro-developmental disorders are based on subjective reports from either patients themselves, their family members, or in the best case scenario, by a clinician [19]. Subjective reports typically exist in the form of multiple choice questions (MCQ) style questionnaires. Some of these questionnaires are administered by a clinician as the patient is interviewed, while others questionnaires are meant to be self-administered by the patient themselves. For example, the Hamilton Rating Scale for Depression (HRSD) [20] is a clinician-administered questionnaire

which is used to provide an indication of depression. Meanwhile, the Patient Health Questionnaire (PHQ-8) [21] is an example of self-administered questionnaire which provides an indication of depression. Similarly, the Young Mania Rating Scale (YMRS) [22] is used to provide an indication of mania severity for individuals with bipolar disorder. It is typically administered by a clinician, although a self-administered version also exists [23].

Regardless of the chosen option, that is, self-administered or clinician-administered, both are prone to bias and, therefore, inconsistencies. Self-assessment based questionnaires essentially rely on individuals to honestly report their symptomatic behaviour. This may not always be true. In fact, for the dataset provided as part of Audio/Visual Emotion Challenge 2017 [13], we note that certain subjects fill in the PHQ-8 questionnaire in such a way that their accumulative score is zero – which may show that they have excellent mental health – but in the interview transcripts these subjects go on to discuss their battles with depression and post-traumatic stress disorder in the past [24]. Meanwhile, Snowden et al. [25] details how behavioural observations by clinicians are also prone to bias – be it intentional or inadvertent – which may compromise their clinical judgement. They argue that differences in clinical training and experience can also adversely affect the homogeneity of clinical opinions, meaning that inconsistencies can occur in clinician-administered questionnaires as well.

According to Bedi et al. [26], psychiatry lacks the capacity to diagnose and subsequently treat mental illnesses due to the absence of objective clinical tests which one finds part of the routine process in other fields of medicine. Furthermore, the lack of systematic and efficient methods to incorporate behavioural cues, which are strong indicators of psychological disorders, is also a major hindrance for the screening process [27]. However, as noted by Solomon et al. [28], despite the inherent flaws of current methods, these methods at least provide a reasonable quantifiable standard to measure the mental and behavioural state of patients. This information can be used for the development of automated screening methods.

1.4 Motivation

As discussed in the previous section, conventional methods for screening of mental and neuro-developmental disorders depend almost entirely on either patients reporting their behaviour or clinicians observing the patients' behaviour as they interact with them. In fact, the lack of systematic ways of incorporating behavioural observations has hampered the capacity of clinicians to diagnose and treat serious mental illnesses [25].

Recent success in human behaviour understanding by virtue of advancements in the field of SSP/AC [4, 5, 17, 29] has led researchers from this community to advocate the development and subsequent use of automated screening

methods. It is reminded here that automated screening methods refer to computational frameworks which can recognise various aspects of human behaviour by processing information conveyed by their social signals such as facial expressions, speech, head motion, and body gestures amongst others.

Automated screening methods have a huge potential to transform healthcare especially for diagnosing serious mental illnesses. Some of the benefits of using automated screening methods in addition to conventional screening methods include objectivity, repeatability, information processing power, and interpretability.

One of the main benefits of automated screening methods is that they can provide objective measurements, which are not affected by subjective bias from clinicians or patients. Another attraction of computerised automation is repeatability. Once the performance of a particular method is found satisfactory, the method can be replicated any given number of times, consistently providing reproducible results. Moreover, advances in the field mean that computer algorithms have the ability to simultaneously track changes in multiple social signals which may be affected by mental and neuro-developmental disorders, something which even clinicians may not be able to do due to human limitations. Finally, these methods can be designed to be interpretable so that they can provide empirical feedback to research domains of psychiatry and psychology.

1.5 Thesis Contributions and Results

In this section, we summarise the contributions of our thesis, which has been organised such that each subsection details the contribution from a chapter.

1.5.1 Limitations and Challenges in Automated Screening

We conducted a literature survey to understand the inherent limitations that exist, and challenges which need to be overcome for developing automated methods. While a discussion is provided in Section 3.5, we summarise key contributions of our survey as follows:

- A major hindrance for research towards the development of automated screening methods for mental and neuro-developmental disorders is limited availability of datasets. This is a result of ethical restrictions which while permitting some researchers to collect and use data, do not permit them to share data with other researchers, mostly due to privacy concerns.

  We appreciate that ethical restrictions exist but also believe that there is a way around this limitation: researchers who have access to private/restricted datasets can compute audio/visual features using standard software such the openSmile toolkit [30], the COVAREP toolkit [31],

and the OpenFace toolkit [32] and share results for their dataset. Others can follow suit. At the end, meta-analysis can be conducted for results obtained from multiple datasets to provide a better understanding of the automated screening task at hand.

In fact, with this in mind, we provide detailed discussion of results for standard feature sets as part of our work on automated screening of Autism Spectrum Disorder in Chapter 6, especially since our dataset cannot be shared publicly due to ethical restrictions.

- Among the datasets which are publicly available (under restrictions of academic research and associated ethics), we note most contain data from a relatively small number of subjects (see Table 3.1 as an example). This makes it difficult to ascertain the efficacy of automated screening methods simply because the trained models are vulnerable to the effects of overfitting.

  There is little one can do when datasets have small number of subjects. Our proposed solution to this limitation is to develop automated screening methods which are inspired from background knowledge of psychology. We surmise that when this approach is followed, proposed automated screening methods are likely to model traits of mental and neuro-developmental disorders than learn datasets by brute-force.

- The limitation due to relatively small size of datasets is further exacerbated by many confounding factors which can affect social signals in addition to possible impairments due to mental and neuro-developmental disorders. This in turn affects the accuracy of automated screening methods. On the basis of our literature survey, we find that the effects of confounding have mostly been ignored in research literature.

- We note that while automated screening methods have great potential to revolutionise mental health care, these methods cannot replace clinicians under the current setup. This is simply because clinicians have access to much more data as well as many years of training and practice.

  For example, clinicians can both, watch and hear, the patient during the interview process. Thus, clinicians have access to information about the patient's facial expressions, head motion, body posture, and speech. In most cases, clinicians will also have access to the patient's medical history, including any medications they take. Meanwhile, datasets currently available for development of automated screening of mental and neuro-developmental disorders only contain information about the patient's face and their speech [10–13, 33].

- As with most machine learning methods, the success of automated screening methods is tied to the quality of data and associated labels. We

note that labels for mental and neuro-developmental disorders are created using either using self-administered [11, 12] or clinician-administered [33] questionnaires, both of which are susceptible to bias and therefore have the potential to introduce label noise.

In light of the surveyed literature, we believe that development of methods for automated screening of mental and neurological disorders can greatly benefit from collaboration between researchers from SSP/AC community and clinicians.

1.5.2 Automated Screening for Depression

Chapter 4 of this thesis is dedicated to the task of automated screening for depression. Here we introduce three new approaches for the task of automated depression screening from audio-visual recordings, namely; turbulence features, craniofacial movement features, and Fisher Vector encoding of spectral low-level descriptors (LLDs). The efficacy of these methods is demonstrated using datasets from 2014, 2016, and 2017 editions of the AVEC challenge on depression screening. We summarise the key contributions of our work as follows:

- We surmise that psychomotor changes due to depression lead to uniqueness in an individual's speech pattern which manifest as sudden and erratic changes in speech feature contours. To this end, we propose a novel set of temporal features, which we called *turbulence features*, to quantify fluctuations in the feature contours of speech features.

  The efficacy of turbulence features was demonstrated as part of our solution for the AVEC 2017 depression severity prediction sub-challenge [13], where we stood 6th overall in the competition, beating the challenge baseline [34]. Amongst various voice quality and prosody features which were part of our investigation, we found turbulence features computed for pitch feature contour to be most useful for the task of automated depression screening.

- We detailed a methodology to quantify specific craniofacial movements, which we hypothesised could be indicative of psychomotor retardation and hence depression.

  The efficacy of these features was tested in terms of the value of Pearson's correlation coefficient [35] with respect to depression severity. We used three sets of recordings from two publicly available datasets from AVEC challenges on depression severity prediction i.e. the AVEC 2014 Depression severity prediction challenge (AVEC 2014 DSC) [12] and the AVEC 2017 Depression severity prediction challenge (AVEC 2017 DSC) [13].

The results demonstrate the efficacy of our proposed craniofacial movement features. Moreover, given that these features are inspired by knowledge of psychomotor retardation from the DSM 5 manual [18], we believe that interpretability of these features will provide meaningful feedback to clinicians for diagnosis of depression.

- We hypothesised that individuals with depression have unique characteristics to their speech spectra. To this end, we introduced Fisher vector encoding [36] of spectral LLDs for quantifying abnormalities within speech spectra of individuals with depression.

  Initially, we demonstrated the efficacy of our proposed approach for the AVEC 2016 Depression classification challenge (AVEC 2016 DCC) dataset [12], where the objective was to identify individuals with and without depression [37]. Later, we extended the idea by adding temporally-piecewise aggregation of Fisher vectors as part of our solution to AVEC 2017 DSC [34]. We beat the challenge baseline whilst using this method.

- We note that datasets released as part of AVEC sub-challenges on automated depression screening (2014, 2016, and 2017 editions) have all used accumulated scores from various depression measurement instruments as labels for audio/visual recordings. The AVEC 2014 DSC dataset [11] used Beck Depression Inventory (BDI-II) [38], whereas the AVEC 2016 DCC dataset [12] and AVEC 2017 DSC dataset [13] used PHQ-8 [21]. While these scores provide a way to quantify depression severity in absence of physical tests for depression [28], there are inherent limitations of using these labels.

  For example, the BDI-II [38] asks patients about feelings of satisfaction, disappointment, and guilt, as well as questions about their weight, appetite, and sex-life. It is obvious that information pertaining to these questions cannot be extracted from audio/visual recordings unless patients are explicitly recorded whilst answering these questions. In that case, one would use speech-to-text conversion and follow it up with natural language processing to learn the answers to these questions.

  This means that under the current set up of datasets, automated screening methods based on audio/visual modalities will continue to have sub-par performance relative to ground truth labels. We argue that it may be worth investing time to devise depression measurement instruments (questionnaires) which specially cater for development of automated screening methods. This would, of course, require significant collaboration between researchers from psychology and SSP/AC.

- In light of our discussion in this chapter and Chapter 3, we believe it is important to emphasise here that while significant inroads have been

made for the task of automated screening of depression, this task is still very much a work in progress.

The most outstanding issue remains the lack of publicly available datasets, which is further exasperated by potentially noisy labels from self-assessment based depression measurement instruments. The many confounding factors such as gender, age, and the nature of speaking tasks means that research for the development of automated methods for screening of depression is likely to continue for at least the near future before these methods are deemed ready for clinical usage.

The work completed for the task of automated screening for depression was published in two conferences, including our participation in the AVEC 2017 challenge.

- Zafi Sherhan Syed, Kirill Sidorov, and David Marshall, "Depression Severity Prediction Based on Biomarkers of Psychomotor Retardation", ACM International Workshop on Audio/Visual Emotion Challenge (AVEC), pp. 37-43, 2017.

- Zafi Sherhan Shah, Kirill Sidorov, and David Marshall, "Psychomotor Cues for Depression Screening", IEEE International Conference on Digital Signal Processing (DSP), pp. 1-5, 2017.

1.5.3 Automated Screening for Bipolar Disorder

Chapter 5 of this thesis is dedicated to the task of automated screening for bipolar disorder. Here we proposed two new approaches for the task of automated screening of bipolar disorder from audio-visual recordings; namely turbulence features and Fisher Vector encoding of Computational Paralinguistics Challenge (ComParE) LLDs. We also introduced Greedy Ensemble of Weighted Extreme Learning Machines (GEWELMs) based classification of these features and demonstrated the efficacy of these methods on both, the development and test partitions. We summarise the contributions of our work as follows:

- Our proposed approaches for automated screening of bipolar disorder from audio/visual modalities is inherently novel for this task since the AVEC 2018 Bipolar Disorder sub-challenge (BDS) [33] provides researchers for the very first time a dataset which contains multi-modal recordings of individuals with bipolar disorder based on structured interviews.

- We surmise that traits of bipolar disorder cause sudden and erratic changes in the contours of social signals and propose *turbulence features* to quantify these changes.

We report that for the task of predicting severity of mania, turbulence features computed for visual modality performed better than the audio modality. In fact, the best result achieved by us on the test partition i.e. UAR = 57.41% uses turbulence features for visual modality. This result exactly matches the best result published as the official baseline, which was a result of fusion of features from audio and visual modalities.

- Fisher Vector encoding of ComParE LLDs achieved best performance in terms of classification accuracy amongst all other features on the development partition. However, their superior performance could not be replicated on the test partition, likely due to overfitting on the development partition. Given limited attempts on the test partition, we could not identify the cause of overfitting on the development partition, although we posit that there are likely to be some confounding factors which influence our machine learning models.

- We investigated the efficacy of four standard feature sets from the openSmile toolkit i.e. Prosody, IS10-Paralinguistics, ComParE functionals, and eGeMAPS features. We found IS10-Paralinguististics and eGeMAPS feature sets to be most useful. It is important to mention here that while organisers report that eGeMAPS features achieve a UAR = 55.03% on the development partition [39], we could not replicate this result, even though we used same experimental settings as reported by them.

- We also investigated whether it is better in terms of accuracy to perform classification over the entire recording as a single entity or to classify each session independently and later perform fusion to yield a label for the recording. Based on our experiments, we report that session based classification is a better option when it comes to classification accuracy. This is somewhat contrary to the findings of Ciftci et al. [33], who reported no advantage of similar segmentation.

- In our attempt of crafting features based on background knowledge of bipolar disorder, we found that certain aspects of behaviour of subjects cannot be probed directly from audio-visual recordings. For example, lack of requirement for sleep is a key behavioural indicator for individuals with mania as per the Young Mania Rating Scale (YMRS) [22, 23]. Now, unless the subject is explicitly asked a question about their sleeping habits, it may not be possible to ascertain how much sleep a particular subject has been having. We attempted to quantify sleepiness using action unit 45 (AU45) [40], which represents blinking, but this approach performed poorly.

  Similarly, sexual activity/interest is another aspect of the YMRS which cannot be directly gauged from audio-visual recordings. For the AVEC 2018 BDS, we found that such questions were not asked by the subjects

in the audio/visual recordings which are provided as part of the dataset. While our aim is to propose features inspired from behavioural characteristics of individuals with mania as per the YMRS, it is necessary to acknowledge the existence of inherent limitations of this approach.

The work completed for the task of automated screening for bipolar disorder has been accepted for the AVEC 2018 workshop co-located with ACM Multimedia Conference, which will take place in October 2018.

- Zafi Sherhan Syed, Kirill Sidorov, and David Marshall, "Automated Screening for Bipolar Disorder from Audio/Visual Modalities", ACM International Workshop on Audio/Visual Emotion Challenge (AVEC), pp. 1-6, 2018.

1.5.4 Automated Screening for Autism Spectrum Disorder

Chapter 5 of this thesis is dedicated to the task of automated screening for Autism Spectrum Disorder (ASD). Our work on the development of automated screening methods for ASD is novel in the sense that we performed numerous experiments and provide discussion on the effects of ASD on speech, in light of research literature, for a previously unpublished dataset. The contributions of our work are listed as follows:

- We manually annotated audio/visual recordings of interview sessions using ELAN software [41, 42] to mark segments of recordings which only contain speech of subjects. This enabled us to undertake investigation into automated screening for ASD using speech.

  While our current work is limited to analysis of speech, our efforts for annotation of these recordings opens up avenues for future research. For example, analysis of facial expressions and body movement when children either speak or are spoken to by the interviewer. Using these annotations, models can also be built to investigate synchrony of dyadic communication between children and interviewer.

  We acknowledge and credit Dr. Catherine Jones (from Cardiff University's School of Psychology) and her team for collecting and providing us these audio/visual recordings. Our work focuses explicitly on social signal processing, not data collection.

- We discussed our feature engineering and classification mechanism for automated screening of ASD from speech for subjects in our dataset. The proposed feature engineering mechanism consists of a four-step process which includes speech segmentation, feature extraction, feature description, and feature selection.

The efficacy of this mechanism was demonstrated in terms of classification accuracy and the identification of highly discriminative speech features using Mann-Whitney U-test based statistical analysis [43] and effect size for statistically significant features [44, 45].

- We investigated the influence of speech segmentation on classification accuracy. To this end we performed experiments using six different segmentation rules. Our results suggest that classification accuracy is dependent on the duration of voiced speech in each speech segment.

- We report based on our experiments that traditional voice quality features such as shimmer, jitter, and HNR are not able to provide discrimination between speech of individuals from TD and ASD groups. In addition, we report that features from the COVAREP voice quality feature set are able to discriminative between the speech of individuals from the two groups.

- We report from experiments conducted using the standard feature sets that subjects with ASD have smaller pitch and loudness variability (in terms of standard deviation), which suggests monotonic speech. This is in line with findings in [46].

- Finally, on the basis of experiments performed in this chapter and in light of published literature, we argue that while the DSM-5 [18] does not currently recognise that the speech production system is affected by ASD, there is enough evidence from research literature as well as our investigation to suggest that ASD may actually have a significant effect on the vocal production system.

1.6 Thesis Outline

The rest of this thesis is organised as follows:

- In **Chapter 2**, we start by discussing the fundamentals of feature extraction from audio, visual, and text modalities. These modalities are commonly used in SSP/AC. We follow this by a brief but directed literature survey for feature engineering mechanisms and machine learning techniques which are used for developing automated screening methods.

- In **Chapter 3**, we start with a brief discussion of psychomotor changes which occur due to mental disorders such as depression and bipolar disorder. These changes can have a profound effect on the social signals of individuals affected by these disorders. Next, we discuss in detail how information conveyed by social signals from audio and visual modalities can be quantified as audio/visual features. As per the scope of the

thesis, we focus on information extracted from the face region from the visual modality and speech from the audio modality. We then discuss in detail inherent limitations which exist and challenges which need to be overcome for the task of developing automated methods for screening mental and neuro-developmental disorders.

- In **Chapter 4** we start with statements of novelty and contributions through our work on development of automated screening of depression. We follow this by describing traits of individuals with depression according to the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) [18]. We also discuss various depression measurement instruments which are used to gauge severity of depression. This provides the foundation for our proposed automated screening methods. Next, we describe datasets used in this thesis and work carried out by other researchers in the field for these datasets. We follow this up with a detailed discussion on our proposed methods, which is supported by experimental analysis and participation in the AVEC 2017 Depression severity prediction challenge. Finally, we end the chapter with a summary of contributions of our work.

- In **Chapter 5** we start with statements of novelty and contributions through our work on development of automated screening of bipolar disorder. We follow this by describing traits of individuals with bipolar disorder according to DSM-5 manual [18]. We also briefly describe states of bipolar disorder as per the Young Mania Rating Scale (YMRS) [22]. This provides the foundation for our proposed automated screening methods. Next, we provide a survey of research literature published for automated screening of bipolar disorder from audio/visual modalities. While limited research literature exists, we were able to identify some features from audio/visual modalities which were previously deemed useful by others. Next, we discuss our proposed methods in detail and provide experimental analysis. We also discuss our submissions for the test partition of the AVEC 2018 Bipolar disorder sub-challenge. Finally, we provide a conclusion based on insights from our work.

- In **Chapter 6** we start with statements of novelty and contributions of our work on the development of automated methods for screening of Autism. We follow this with a literature survey to identify speech features which have high discriminative power when it comes to identifying individuals with ASD. We then describe the dataset which we use in our work. This is followed by a discussion on a feature engineering and classification mechanism to screen for ASD using speech features. We then provide a discussion on the efficacy of standard feature sets for the task at hand. Finally, we end the chapter with a summary of contributions of our work.

- Finally, in **Chapter 7**, we conclude this thesis by covering the achievements of this work. We also discuss possible directions for continuation of research conducted as part of our effort.

Chapter 2

Fundamentals of Automated Screening Methods

2.1 Introduction

The purpose of automated screening methods is to develop computerised methods which can identify groups of individuals which are different from each other with respect to a particular trait. In context of this thesis, our aim is to develop automated screening methods which can differentiate between individuals who have mental and neurodevelopment disorders and healthy individuals. The term *automated screening method* in the domain of Social Signal Processing/Affective Computing (SSP/AC) refers to a computational framework that can recognise various aspects of human behaviour by processing information conveyed by their social signals [3, 6].

As illustrated in Figure 2.1, the process flow for the development of automated screening methods consists of three fundamental steps. The first step is to identify relevant social signals for the task at hand. For example, if the task is to recognise emotions of subjects from telephone recordings then speech is the relevant social signal. On the other hand, if the task is to recognise emotions from twitter posts then relevant social signals will be from the text modality. Information contained in these social signals needs to be represented as a measurable quantity before it can be processed by machine learning algorithms. This representation is called a *feature*. Features computed directly from social signals are called low level descriptors (LLDs). While it will be clearer in subsequent sections (see Section 2.3), these features are 'raw', and typically require further processing before they can be passed on to machine learning algorithms.

The next step in the pipeline is called feature engineering. As the name suggests, this step involves further processing of LLD features into a form which is most suitable for machine learning algorithms. This process is typically governed by domain knowledge of the social signals, and the features resulting

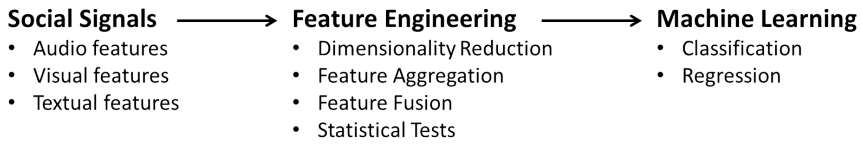


Figure 2.1: Workflow diagram for automated screening methods in social signal processing

from these processing steps are called *intermediate features* or *higher level descriptors (HLDs)*. There are various sub-steps involved in the task of feature engineering. These include feature aggregation, dimensionality reduction, and feature fusion. Furthermore, statistical tests can be used to determine the discriminative power, and therefore usability of intermediate features before they are finally passed down to machine learning algorithms.

The final part of the automated screening pipeline is to apply suitable machine learning algorithms. The most fundamental decision, that is the choice between regression and classification, is dependent on the task at hand. For example, if the target variable for automated screening task is continuous then machine learning algorithms which specialise at regression are selected. On the other hand, a suitable classifier needs to be selected if the target variable is discrete. It needs to be mentioned here that feature engineering and machine learning steps typically overlap, and are tuned until either a desired or an acceptable performance is achieved for the task at hand.

Our aim in this chapter is to provide a gentle introduction to the fundamentals of automated screening methods as used in the field of SSP/AC. We do not explicitly focus on screening for any particular behavioural characteristic in this chapter, instead provide a more general discussion. To that end, we provide a more targeted discussion for development of automated screening methods for mental and neuro-developmental disorder in Chapter 3.

The rest of this chapter is organised as follows: we start with discussing fundamentals of feature extraction from audio, visual, and text modalities. These modalities are commonly used in SSP/AC. We follow this by a brief but directed literature survey for feature engineering mechanisms and machine learning techniques which are used for developing automated screening methods.

2.2 Audio, Visual, and Textual Modalities

Human beings exhibit social signals through which they inform and communicate their emotional behaviour and mental state [4]. While there are a large number of social signals, the ones which are most commonly used for automated screening of mental and neuro-developmental disorders are those

which convey information about the face and speech (including its linguistic content), thus requiring extraction of features from three modalities i.e. audio, visual, and text.

It is important to mention here that while other modalities may aid the task of automated screening, one is restricted to modalities which are available in datasets. For example, the audio modality is confined to speech segments and ambient noise in between speech segments. The visual modality is confined to facial information i.e. most publicly available datasets simply do not provide information about the subjects' hand gestures, torso or leg movement. While initially not considered particularly important, textual modality has recently become relevant, especially when video recordings contain structured interviews.

2.2.1 Audio Features

Speech is arguably the most important modality for studying behavioural cues of an individual. It is a measurable quantity that can be used as a correlate for various emotional and mental states of an individual. The use of speech as a behavioural cue is motivated by several neuro-physiological studies, as detailed in [47] and references therein, which report that regions of brain responsible for social functioning also affect speech production.

The speech modality has been widely used for the purpose of emotion recognition [48], where the objective is to infer feelings of anger, happiness, surprise, disgust, sadness, and fear [7], as well as arousal, valence, and dominance [11]. Speech has also been used to screen for various mental, neuro-developmental, and behavioural disorders such as cognitive load [47], depression [11, 12], schizophrenia [49], psychosis [26], Parkinson's disease [50], Alzheimer's disease [51, 52], and ASD [53] amongst many more.

In order to use speech in a computational framework, one needs to compute representative features from recordings of speech using appropriate digital signal processing (DSP) algorithms. The non-stationary nature of speech mandates that DSP algorithms process speech as short time duration chunks of speech (called frames), rather than processing the entire speech recording as a single block [54, 55]. These frames are typically between 15–40 ms in duration and it is assumed that over this duration, the speech signal is quasi-stationary, thus enabling the application of DSP algorithms. Since these features are computed over such a small temporal resolution, they are called LLDs. It is important to mention here that these LLDs only provide *local* information about the speech recording, and extra processing is required for a global representation of the speech recording, as discussed in Section 2.3.2.

A large number of features can be computed from the speech signal. In order to use these features in an informed manner, taxonomising them is necessary. Defining taxonomy is, however, not trivial since one can define multiple taxonomies depending on which principles are used. Speech LLDs

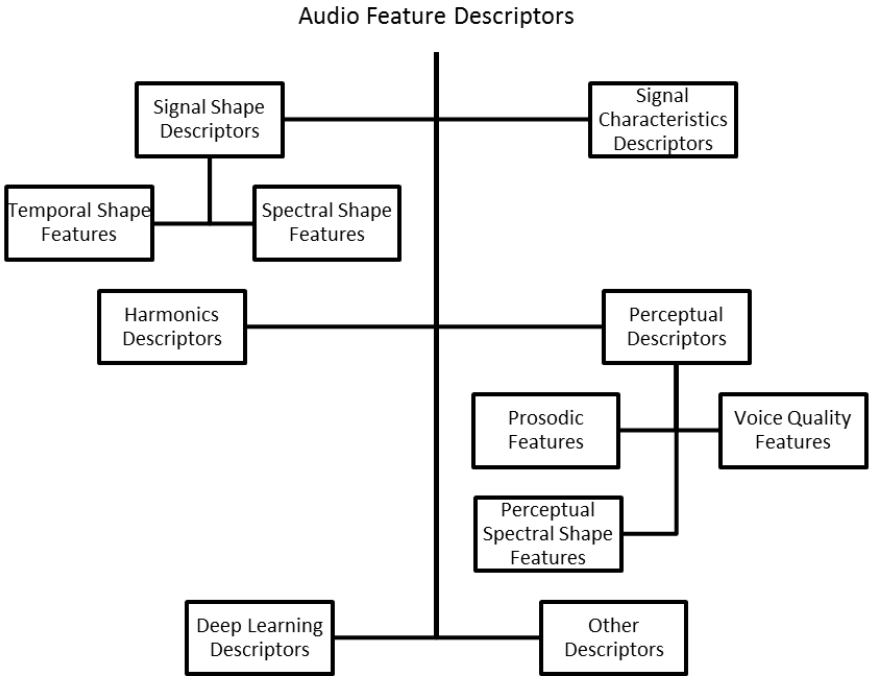


Figure 2.2: A taxonomy of audio descriptors for social signal processing

can be categorised as either temporal, spectral, cepstral, or time-frequency in nature, on the basis of the domain of computation. On the basis of their dimensionality, speech descriptors can be categorised as single dimensional (for example, pitch and energy) or multi-dimensional (for example, Mel-Frequency Cepstral Coefficients). Speech descriptors can also be categorised according to their psychoacoustic characteristics, where descriptors are computed on the basis of models of the human hearing process. Perceptual features can further be categorised into prosodic features, voice quality features, and perceptual spectral/cepstral features.

Mitrovic et al. [56–58] have advocated various categorisation methods of audio features, and motivated by their work, we propose categorisation which we believe is best suited from a purely social signal processing perspective. Our proposed categorisation principles are quite straightforward, as illustrated in Figure 2.2, where we prioritise the categorisation of speech features on the basis of their perceptual meaning, and if that is not possible, we categorise them according to the physical domain of computation and the characteristics of speech signal.

Signal Shape Descriptors

This category includes features which describe the temporal and spectral shape characteristics of the speech signal. Temporal shape features provide information about the shape of the speech signal in time-domain i.e. characteristics of its amplitude over time. Examples include rate of signal increase/decrease, duration of signal increase/decrease etc. Meanwhile, spectral shape features provide information about the shape of magnitude spectra of the speech signal. For example: spectral centroid, spectral spread, spectral slope and roll-off frequency etc.

Signal Characteristics Descriptors

These descriptors provide information about the speech signal which does not necessarily has to be its shape. For example, features such as peak energy, root-mean-square energy, autocorrelation width, zero crossing rate etc. belong to this category.

Harmonic Descriptors

This category includes features which provide information about harmonic content of the speech signal and the concentration of energy around those harmonics. A new category is required for these features since they explicitly focus on specific frequencies rather than the entire speech spectra. Examples of harmonic descriptors include: formant frequencies, harmonic-to-noise ratio, cepstral peak prominence etc. This category also includes features which provide information about the relative energy at different frequencies. For example, the energy difference between first and second harmonic, commonly represented as $H_1 - H_2$, energy difference between the first harmonic and the highest harmonic in the range of the third formant i.e. $H_1 - A_3$.

Perceptual Descriptors

Features in this category are computed using signal processing models inspired from the human speaking and hearing process. Perceptual descriptors can include three types of features, i.e. (a) prosodic features, (b) voice quality features, and (c) perceptual spectral shape features. Prosodic features provide information about the patterns of rhythm in the speech signal. For example, pitch (as represented by the fundamental frequency), intensity, jitter, shimmer etc. Voice quality features provide information about the voice quality, examples include normalised amplitude quotient (NAQ) and quasi-open quotient (QOQ) are known to provide information about voice tenseness, cepstral peak prominence (CPP) provides information about the amount of breathiness in speech. And, perceptual spectral shape features are essentially similar to spectral shape feature, but the difference is that the magnitude-spectra

is warped to match psychoacoustic frequency scales such as Mel and Bark. Examples include Mel frequency cepstral components (MFCCs), perceptual linear prediction coefficients (PLPs) etc.

Deep Learnt Descriptors

Given the recent success of deep learning methods, especially in the field of computer vision, a large number of researchers have proposed using deep learning to learn features from the speech signal. There are two fundamental approaches for using deep learning on the speech signal. The first and most commonly used approach is to compute features from the speech spectrogram. This approach feeds speech as two-dimensional images to any type of deep neural network which can accept images as input [59]. The second approach is to use speech signal as a one-dimensional signal. While not as common as first approach, deep learning architectures such as Wavenet from Google's DeepMind [60] use this approach.

Other Descriptors

In principle, any mathematical equation with an input and output is a DSP algorithm which when applied to the speech signal will compute some features. In the *other descriptors* category, we propose to include any feature which does not suit other categories, such as features computed using wavelet transform [55] or the fractional Fourier transform [61].

A number of audio feature descriptors have been proposed in research literature with the aim of standardising the use of these features. Most recent publications have, in fact, focussed on using standard feature sets. The benefit of using standardised feature sets is that it helps in comparing the efficacy of specific speech features for various applications. The most popular feature sets include the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) and its extended version, the eGeMAPS [58], the Cooperative Voice Analysis Repository for Speech Technologies (COVAREP) toolbox [62], the VoiceSauce toolbox [63], the AVEC 2013/2014 speech feature set [10,11], the Computational Paralinguistics Challenge (ComParE) feature set [64], and the Audio Analysis Library toolbox [65].

2.2.2 Visual Features

The face conveys a multitude of information about an individual such as their age, gender, social background, and in most cases what they are thinking, feeling and what they intend to do next. The validity of head pose, head nods, and facial expressions as signs to study human behaviour are already well established in research literature from the field of psychology [66, 67]. In fact, several mental disorders are manifested as an altered facial activity.

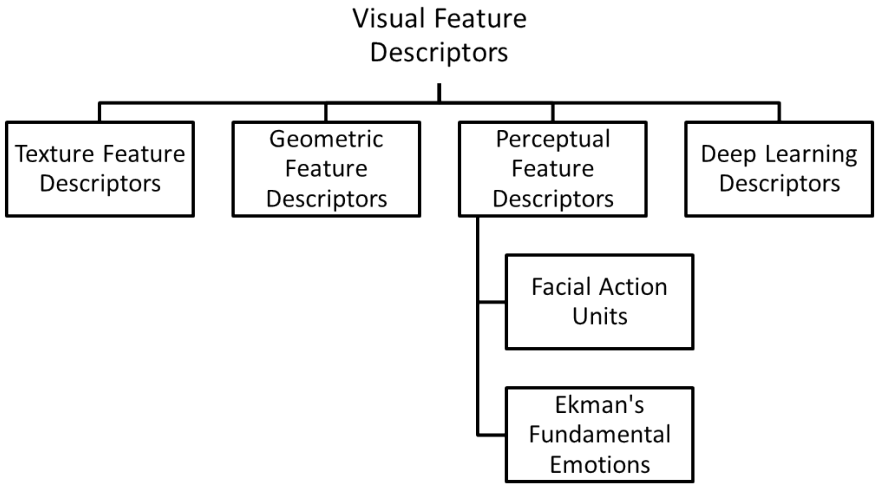


Figure 2.3: A taxonomy of visual descriptors in social signal processing

This has motivated researchers from the SSP/AC community to work on automated quantification of facial observations for psychiatric diagnosis of autism [68–70], depression [71], schizophrenia [72, 73], and psychosis [74], just to name a few. However, this work is still very much under development.

Facial observations can be quantified using a myriad of visual features. Following the precedent of taxonomising audio features, we propose to categorise visual features as illustrated in Figure 2.3. This taxonomy prioritises allocating features on the basis of their perceptual meaning in terms of human emotions [2,7]. If a visual feature descriptor does not directly have a perceptual meaning, it is categorised according to the method of computation.

Texture Feature Descriptors

Facial muscle movement, which generates facial expressions, manifests as changes in skin texture with appearance of facial furrows and wrinkles. It follows that one can quantify facial muscle movement by crafting visual features which focus on the appearance (texture) of the face region. The use of texture based feature descriptors requires pre-processing of the video on a frame-by-frame basis [75]. The first step is use an appropriate face detection algorithm, such as the Viola-Jones method [76], to find the region of the image which contains the face so that it can be cropped out for further processing. This step is required because the image may contain irrelevant information i.e. objects other than the face. This step is followed by a registration process, which transforms faces from all frames in the video to pre-defined, fixed number of pixels.

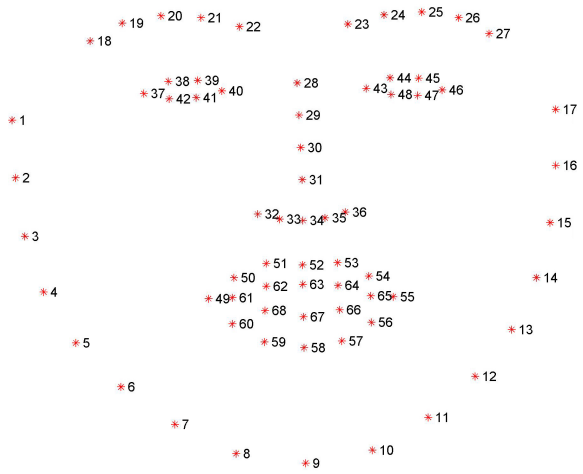


Figure 2.4: Reference for numbering of 68 point facial landmarks.

Once the images are registered, one can start the actual computation of texture features. The simplest texture based feature is raw pixel intensity but is highly susceptible to illumination and skin tone changes, but better alternatives are available. For face studies, the most common texture features include the Local Binary Pattern (LBP) [77], Histogram of Oriented Gradients (HOG) [78, 79], and Local Phase Quantisation (LPQ) [80] just to name a few. While legacy texture feature descriptors only provide visual information of each frame independently, newer methods offer the ability to measure intra-frame dynamics of texture by leveraging the three orthogonal planes (TOP) concept [81]. The TOP computes features across three planes i.e. the spatial plane (height and width) and the two temporal planes (height-time and width-time).

Geometric Feature Descriptors

Geometric features are computed from facial landmarks — a term used to describe contour points of key positions on the face, such as eyes, the nose, the nostril corners, the mouth, the eyebrows, and the chin, as illustrated in Figure 2.4.

As the name suggests, geometric features are measures of the shape, size, and relative position of all or a subset of these facial landmarks. The most commonly used geometric feature is the measure of the distance between facial landmarks. Angles between various facial landmarks have also been used as features and so has the area enclosed within regions covered by facial landmarks. In principle, any geometric measure can be computed for facial landmarks and be called a geometric feature descriptor for the face.

Geometric features require that facial landmarks be already computed for each frame of the video recording, which can be done using one of the many publically available libraries such as dlib [82], OpenFace [32], and Chehra

tracker [83]. Facial landmarks typically contain geometric variations (such as scale and rotation) as a result of an individual's head-pose with respect to the camera recording the video. While it is common practice to remove these variations through a registration process [75], doing so in some cases may compromise the quality information provided by the facial landmarks. For example, while face registration is useful for recognising facial expressions, this process is detrimental if the objective is based on studying head-pose.

Perceptual Feature Descriptors

This category of visual feature descriptors has perceptual meaning, typically based on the theory of facial expressions and emotional behaviour. We propose to sub-categorise perceptual visual features as: (a) Facial Action Units, and (b) Ekman's Fundamental Emotions. This sub-categorisation is strictly based on the output available from popular software tools such as OpenFace [32], the Computer Expression Recognition Toolbox (CERT) [84], and IntraFace [85], which are in turn motivated from research work in Computer Vision to automatically recognise facial action units and emotions. While it is possible to design and develop machine learning methods to both facial action units and fundamental emotions from scratch, using of off-the-shelf tools enables replicability and objective comparison of empirical research in social signal processing.

Facial Action Units, simply called Action Units or AUs and sometimes FAUs, are a set of rules which describe the movement (actions) of individual muscles or groups of muscles on the face. AUs are based on the facial action coding system (FACS) [1, 2], which is the de facto standard to encode facial muscle movement [75, 86]. As examples, consider AU4 which is the Brow Lowerer and represents the Depressor Glabellae, Depressor Supercilli, Currugator facial muscle, and the AU12 which is the Lip Corner Puller and represents the Zygomatic Major facial muscle. There are 32 FAUs in total according to the revision made to FACS in 2012 [2]. A summary of most common AUs has been provided in Table 2.1, although, we would direct the reader is directed to the visual guide provided in [40] for further details on AUs.

Perceptual visual features can also be categorised on the basis of which emotions they represent. Paul Ekman argued in [7] that there exist six basic emotions which are hard wired in human beings. The emotions include anger, disgust, fear, happiness, sadness, and surprise. Furthermore, Ekman argued that these emotions can be decoded as manifestation of facial action units — an idea which has already been corroborated by many years of research studying facial expressions [86]. For example, the emotion of happiness can be represented as presence of AU6, AU12 and AU25, meanwhile the emotion of sadness can be represented as presence of AU1, AU4, AU6, AU11, AU15, and AU17. In order to have a perceptual understanding of actions units and emotions, the reader is encouraged to visit the FACS video guidebook [40].

Table 2.1: Summary of most common Action Units as per [1, 2]

AU	Description	AU	Description
1	Inner Brow Raiser	14	Dimpler
2	Outer Brow Raiser	15	Lip Corner Depressor
4	Brow Lowerer	17	Chin Raiser
5	Upper Lid Raiser	20	Lip Stretcher
6	Cheek Raiser	23	Lip Tightener
7	Lid Tightener	25	Lips part
9	Nose Wrinkler	26	Jaw Drop
10	Upper Lip Raiser	28	Lip Suck

Deep Learnt Features

As with most domains, deep learning can also be used to compute visual features from facial images. One possible approach is to use pre-trained models for object detection such as AlexNet [87], VGGNet [88], GoogLeNet [89], and ResNet [90] amongst many more to extract features directly from face images. Another approach can be the use of models specifically trained on faces such as FaceNet [91] to extract features. It is also possible to build deep learning models from scratch to extract features from texture, geometric and perceptual features already extracted, as implemented by Yang et al. [92].

2.2.3 Textual Features

While being very important for depression screening from blogs and tweets in the domain of Natural Language Processing (NLP), the text modality has largely been ignored in social signal processing due to the fact that datasets, until recently, did not contain textual data.

This changed with the AVEC 2016 dataset [12] which contains data about individuals being interviewed by a virtual agent. Apart from audio and visual modalities, this dataset also contains transcripts of the communication between the individual and the virtual agent. Nevertheless, the scope of our work is restricted to the audio and visual modality only, and the reader is referred to [93, 94] for further details on the textual feature for automated screening of mental health illnesses.

2.3 Feature Engineering and Machine Learning

In this section, we discuss various features engineering mechanisms and machine learning methods which form key parts of the computational framework for the recognition of mental and neuro-developmental disorders.

2.3.1 Cross-Validation

Cross-validation (CV) is a technique which provides a measure of the ability of a machine learning algorithm to generalise to previously unseen data. The principle behind CV is to partition training partition into two independent subsets called the *training subset* and *development subset*. The training subset is used for training the machine learning algorithm, rather than the original training partition, while the development subset is used for testing purposes, rather than the original test partition. The performance of the algorithm on the development subset provides a way to gauge the generalisation ability of the model. CV is advantageous because it helps in fine-tuning the model without having to use the test partition. For further details on CV, the reader is referred to [35].

The most basic type of CV involves partitioning the original training partition in such a way that the training subset contains all but one example, while the development subset contains the remaining example. This approach is called leave-one-out CV (LOOCV), and is commonly used when the number of examples in the main training partition is small, which is often the case in SSP/AC as discussed later in Section 3.5. A major drawback of this approach is that the testing process needs to repeat individually for each example of the original training partition, which can be time consuming. However, there is little to no choice but to use LOOCV for measuring generalisation ability of the model when the size of the data set is small.

K-fold cross validation is another version of CV, which is also commonly used in SSP/AC. In this approach the original training partition is partitioned into K equal sized subsets. Here, $K-1$ subsets are used for training and the last remaining subset is used for testing purpose. LOOCV can, in fact, be considered a special case of K-fold CV where K is equal to the number of training examples.

2.3.2 Feature Aggregation

Feature Aggregation is an approach through which low level feature descriptors are summarised to create features which provide global information about audio/visual recordings.

The audio and visual LLDs described in Section 2.2.1 and Section 2.2.2, respectively, are computed on a frame-by-frame basis, thus providing only local information. In order to create a global representation of these recordings using LLDs, one needs to aggregate information provided by these features. There are a number of approaches to feature aggregation including (a) functionals, (b) Gaussian Mixture Models (GMM), (c) Bag-of-Words (BOW), and (d) Fisher Vector encoding.

Functionals

This method uses descriptive statistics for aggregation. Here, functionals of descriptive statistics such as mean, variance, maximum, median etc. are used to summarise sequences of LLDs. In the simplest case, a scalar feature can be created by applying one of the functionals to LLDs. For example, using maximum functional on the pitch contour provides information about the largest pitch of an individual during a speech recording. In most cases, however, a single functional is not enough to adequately provide global representation of the recording, and one needs to consider multiple functionals.

Gaussian Mixture Models

It is also possible to use Gaussian mixture models (GMM) [95] for feature aggregation. GMM provides a probabilistic model for representing the sequence of LLDs in terms of the parameters of Gaussian functions. The motivation for using GMMs is that it provides a fuzzy alternative to the functional approach. When GMMs are used for summarisation, a super-vector is used as the feature provided as input to the machine learning algorithm. This super-vector can include all or a combination of the means, variance/covariance, and the mixture weights of the GMM [96].

Bag-of-Words

While the GMM based approach summarises LLDs using a probabilistic model, the bag-of-words (BOW) approach does so by using histogram of *words* [97]. In the context of BOW for audio-visual features, the term words refers to a code-books which are created by clustering LLDs from each of these domains. The BOW feature then is a histogram which describes the frequency of occurrence of each word.

Fisher Vector encoding

This approach combines the advantages of both generative and discriminative approaches for machine learning [98]. The process flow for Fisher Vector encoding starts with building a generative model (typically, using GMM) of LLDs, and later computing the Fisher kernel from this generative model. Essentially, FV measures the deviation of the LLDs from the generative model. Fisher Vectors are quantified using first and second order statistics of the gradient of the sample log-likelihood with respect to the model parameters [36, 99].

2.3.3 Dimensionality Reduction

While the feature aggregation methods discussed in previous section provide aggregation from a sequence of LLDs to a fixed length feature vector, it is often

the case that the length of this vector is large. In machine learning terminology, the length of the feature vector is called its dimensionality. It is often the case that when the number of training examples is smaller than or equal to the dimensionality of the feature matrix, machine learning algorithms fail to learn a meaningful representation from data and efforts need to be taken to reduce the dimensionality of the feature matrix [100, 101]. There are various methods for dimensionality reduction, which can be divided into two categories: (a) Feature Projection and (b) Feature Selection.

Feature Projection

In this approach features are projected from a high-dimensional space onto a lower-dimensional space, ideally preserving or even increasing the separability of each feature. Feature vectors from all training examples are concatenated as a matrix and the feature matrix is used to learn the mapping process. This approach for dimensionality reduction is very popular and mostly used in the form of principal component analysis (PCA), while other methods such as canonical correlation analysis (CCA) [102] and partial least squares (PLS) [103] have also found applications, depending on the pattern recognition task at hand [101].

It is also possible to use feature summarisation methods for dimensionality reduction. For example, a GMM with smaller number of Gaussians can be used to reduce the dimensionality of a GMM super-vector with a larger number of Gaussians, as used by Alghowinem et al. [104].

The major drawback of feature projection methods for dimensionality reduction is that the transformation from high to low-dimensionality removes the notion of interpretability of features i.e. features in the low-dimensional space cannot be interpreted easily even if they were interpretable on the high-dimensional space. This brings us to the second option for dimensionality reduction.

Feature Selection

As the name suggests, this approach is based on selecting features from the high-dimensional feature vector which contribute to improve the performance of the machine learning task at hand — be it regression or classification. The main benefit of feature selection methods is that only meaningful features are retained, although this comes at the risk of losing the ability of the model to generalise. The reader is referred to the excellent tutorial by Guyon et al. [100] for a discussion on the principles of feature selection and to [105] for a more recent review on feature selection methods.

There are three subcategories of feature selection methods i.e. (a) filter based approach, (b) wrapper based approach, and (c) embedded approach. Filter based feature selection performs feature selection independent of the clas-

sification or regression algorithm being used for the selected features. However, filter based methods do depend on whether the pattern recognition task at hand is classification or regression. For example, filter based feature selection methods for classification include t-test, Mann-Whitney U-test and minimum redundancy maximum relevance (mRMR) [106], meanwhile for regression Pearson correlation coefficient, Spearman correlation coefficient, Relief [107] algorithms can be used for regression. The second category of feature selection methods is called the wrapper based approach. In this method, the classification or regression algorithm plays a key role. The idea is to incrementally add or remove a feature or set of features and monitor the performance of the pattern recognition algorithm. The process is iteratively repeated until the performance no longer improves. Examples of wrapper based methods include sequential forward search (SFS) and sequential backward search (SBS). Embedded feature selection approach performs the task of feature selection along with classification or regression in one go. Most prominent examples of embedded feature selection methods include LASSO, Ridge regression, elastic-net and partial least squares regression (PLSR), just to name a few [100, 101].

2.3.4 Feature Fusion

Feature fusion is a method to combine information from various features with the aim to improve the performance of the machine learning system, for example the results of classification or regression based on a particular metric [108]. Feature fusion can be implemented in two fundamental ways i.e. (a) feature level fusion, and (b) decision level fusion.

Feature level fusion is quite trivial. In this approach multiple features are simply concatenated with each other, thereby creating a (larger) feature vector. Meanwhile, decision level fusion offers an alternate approach, which takes place once the classifier takes a decision on the label of the input feature vector.

Decision level fusion can be implemented in two ways i.e. by using either hard decision or soft decision. Hard decision is usually implemented by majority voting on the labels predicted by the classifier using logical AND or OR operations. Soft decision fusion typically takes place on the probabilistic outputs of the classifier, or by combining non-probabilistic outputs using suitable weights.

Feature level fusion typically has a poorer performance when compared with decision level fusion. This is understandable since it increases the dimensionality of the feature vectors and inherently makes the machine learning task more complex. Classifiers and regressors may struggle to learn meaningful representation due the *curse of dimensionality* problem and yield relatively poor performance. For the task of automated depression screening, the authors of [109,110] report that feature level fusion led to poorer performance compared to decision fusion.

While decision level fusion generally performs better than feature level fusion, it has certain limitations. For example, there is no absolute rule governing whether hard decision fusion should use majority combining or use Boolean AND or OR operations for feature fusion. In fact, it is often required that one probe multiple operations and select the best performing one, as evident in the work of [111]. Meanwhile, soft decision fusion requires optimisation of weight parameters which are required to combine the classifier outputs. For example, the authors of [112] assign the weights randomly and optimise them using cross-fold validation and those in [113] train a Kalman filter [114] for this purpose. This not only increases the computational cost but also makes the feature fusion process vulnerable to overfitting.

2.3.5 Classification and Regression

Various types of classification and regression algorithms have found use in the computational framework for automated screening of human behaviour. These algorithms include support vector machine (SVM) [115–117], random forest (RF) [118, 119], decision trees [120], relevance vector machine (RVM) [121], extreme learning machine (ELM) [122, 123], and logistic regression [124].

We note that SVM has been the most popular algorithm, finding application in tasks related to automated recognition of depression [102, 125–131], emotions [132–135], public speaking ability [136–139], ASD [69, 70, 140–145], Schizophrenia [73], and paralinguistic activity [64,146–148] amongst many more. For the task of automated screening of depression, decision trees [111, 149–151] and random forest [28,92,131,152,153] have also been used. Amongst other classification and regression algorithms, RVMs have been used in [154–159], ELMs in [128,160,161], and logistic regression has found use in [27,39,53,155,162–168].

Our literature survey suggests that there is no particular method for selecting a classifier/regressor for development of automated screening methods, although SVM does appear to be the 'go to' algorithm. This is likely because classification/regression tasks related to the automated screening of human behaviour are inherently complex. For example, classes are rarely separable and it is typical to have a large overlap between the classes due to the nature of how labels are assigned (we visit this aspect of SSP/AC in Section 3.5.4). Kachele et al. [113], therefore, advocate the use of an ensemble of classifiers/regressors for tasks related to recognition of affective states and depression. They believe that since classes are not linearly separable, a single classifier would have poor performance as it tries to learn contradicting data points or simply creates a best fit of features which may not represent the actual learning task. Another aspect which affects the choice of classification/regression algorithm is the amount of available data. Williamson et al. [128] argue that with limited data, least squares based approach would perform better than stochastic gradient descent approach of SVM. In fact, this motivated them to use ELM instead of SVM. Scherer et al. [126] report that a simple linear discriminant classifier [120]

performed better than SVM for the task of depression recognition, however, they surmise that SVM could have provided a better result had their grid search been wider.

In subsequent sections, we provide a brief introduction of classification and regression algorithm which we have used in this thesis. These include SVM, ELM, and partial least squares regression (PLSR).

Support Vector Machine

The SVM was introduced by Cortes et al. [116] as a linear classifier for binary classification tasks. It models the decision boundary between two classes as a separating hyperplane. Extensions by Drucker et al. [117] enables SVM to also be used for regression tasks, commonly known as support vector regression (SVR). While the legacy SVM was a linear classifier, it can be extended to perform classification/regression when separability between classes in non-linear. This is achieved by first projecting feature matrix into a hyperspace using a suitable kernel such that linear separability exists in that hyperspace, before SVM is used. The task of finding a suitable kernel is not trivial, and typically requires cross-validation for tuning parameters of the various kernel functions. Commonly used kernel functions include linear kernel, Gaussian kernel, and polynomial kernel [169].

Given that there are a number of options for classifiers/regressors, it has become common to use SVM for reporting baseline classification/regression results. For example, Interspeech ComParE challenges provided baseline classification results using the SVM with a linear kernel [146, 148] whilst using implementation of SVM provided by the WEKA toolkit [170]. Meanwhile, the recently concluded AVEC 2018 sub-challenge on Bipolar disorder [39] used the LIBLINEAR toolkit [171] implementation of linear SVM.

Extreme Learning Machine

ELM was introduced by Huang et al. [122, 123] as single layer feed-forward neural network where the hidden layer is assigned randomly generated weights. These weights are not updated during the training process, instead, the classifier works by mapping the output from the hidden layer to the training labels using a least squares fit.

While Huang et al. argue that even with random weights, the hidden layer can learn useful representation of input data which can be exploited by designing a suitable output layer, we do not find their argument to be absolutely true. In fact, as part of our experiments with ELM we found the classification performance with ELM is highly dependent on the choice of random weights i.e. the *seed* used to generate random matrices (see Section 5.6.4 for details).

However, we do agree with Huang et al. that compared with other classification algorithms, ELM enjoy an outstanding advantage of very fast training

times, which enables one to undertake a grid search for the best random weights for the hidden layer.

Partial Least Squares Regression

PLSR is a regression algorithm which is most useful when the feature matrix has collinearities and in cases when the dimensionality of feature matrix is smaller than the number of training examples [103, 172]. PLSR involves two fundamental steps. In the first step, the original feature matrix is transformed such that the new matrix contains uncorrelated components instead of original features. In the second step, least squares regression is applied to map the new feature matrix with the target variable. The PLSR only has one tuning parameter i.e. the number of components to retain, which can be optimised using an appropriate metric (such as the mean square error) on the development partition. Typically, the performance of PLSR will improve as the number of components is increased, however, this also makes PLSR prone to overfitting.

2.3.6 Statistical tests

Statistical tests are commonly used for the development of automated screening methods in order to provide insights and inferences about features and their association to the target variable. We note from our literature survey that two fundamental types of statistical tests have been used for the development of automated screening methods. These include tests to determine the outcome of a hypothesis and the second task is to determine the strength of correlation

In the following sections, we provide a brief introduction to hypothesis tests and correlation tests, in context of their application in SSP/AC.

Hypothesis Tests

The t-test and Mann-Whitney U-test fall under the category of statistical hypothesis tests. These tests are used to determine if the data (which in our case are audio, visual, or text features) from two groups (which in our case are labels, such as depressed and not depressed) is significantly different from each other. While similar in application, there are subtle differences between the t-test and the U-test [35]. For example, the t-test can only be used under the assumption that data points for both groups are normally distributed, whereas no such assumption is need for the U-test. Moreover, the t-test quantifies the difference between mean values of the two groups, whereas the U-test quantifies the difference between median values of the two groups.

The hypothesis that whether the difference between the two groups is statistically significant is determined through the probability value of these tests. The probability value, commonly called p-value, indicates how likely it is to observe a difference between the groups whilst using these tests due to noisy observations when is in reality no difference between the groups [35].

For example, a small p-value means that it is unlikely that the difference in mean/median values of the two groups is due to noisy observations and that the data from the two groups is indeed different. Meanwhile a large p-value means that the difference between the mean/median values is likely to have occurred due to noisy observations.

In SSP/AC, the t-test test has been used in applications pertaining to depression [28, 52, 104, 125, 157, 173–175], schizophrenia [72, 73], bipolar disorder [176], psychological distress [177], and public speaking ability [137]. It is important to mention here that Alghowinem et al. [104, 174, 175] have used t-tests for feature selection instead of hypothesis testing. Meanwhile, the U-test has also been used for recognition of depression [178–180], ASD [145, 181, 182], bipolar disorder [183,184], and cognitive impairments [185]. Amongst these, we find that $p = 0.05$ is typically chosen as the threshold for determining whether or not a statistically significant different exists between the two groups.

Sullivan and Feinn [44] argue emphatically in favour of effect size, which represents the magnitude of the difference between groups, to report statistical analysis to support p-values, and quote Jacob Cohen (a statistician, infamous for his work on statistical power and effect size) as '*the primary product of a research inquiry is one or more measures of effect size, not P values*'.

We find motivation to reporting effect-size for statistical analyses from Fritz et al. [45], who argue that reporting effect sizes not only allows the comparison of effects in a single study, but also permits meta-analyses across studies. While not previously common, we note a growing trend where the effect size based on the Hedge's g [45] has been reported for SSP/AC application related to depression [125, 126, 173, 186–189], public speaking ability [136, 137, 190, 191], and ASD [145].

Correlation Tests

Correlation tests are used in statistics to measure the strength of relationship between two variables. In context of SSP/AC, these are used to determine the relationship between an audio, visual, or text features and the continuous-valued target variable, which, for example, could be the depression severity.

The two most commonly used correlation tests are the Pearson Correlation Coefficient (Pearson r) and Spearman Correlation Coefficient (Spearman ρ) [35]. Similar to the t-test and U-test, there are subtle differences between Pearson r and Spearman ρ. The Pearson correlation coefficient is used to evaluate relationship between two continuous variables under the assumption that the relationship between the two variables is linear and that the both are normally distributed. Meanwhile, the Spearman ρ is used to evaluate monotonic relationship between two continuous or ordinal variables. Unlike, Pearson r, Spearman ρ does not have restrictions for the normal distribution of the two variables. The correlation coefficient r/ρ can take values between -1 and 1. A value of 1 indicates perfect correlation, whereas a value of -1 indicates perfect

but negative correlation. A correlation coefficient value of 0 means that no correlation exists.

We find that Pearson's Correlation Coefficient has found use in applications related to screening of depression [125–128, 130, 178, 192, 193], ASD [194, 195], bipolar disorder [33, 183, 184], public speaking ability [136, 137, 190, 191], schizophrenia [73], and psychosis [196]. Spearman's Correlation Coefficient has also found use in various applications related to screening of depression [127, 197, 198], psychosis [196], ASD [194, 195], and public speaking ability [190].

2.4 Summary

In this chapter we provided a gentle introduction to the fundamentals of automated screening methods as used in SSP/AC. We started with a discussion on the three most popular modalities as used in SSP/AC [4] i.e. audio, visual, and text.

We followed this by a discussion on various feature engineering mechanisms including feature aggregation based on functionals, GMMs, BoWs, and Fisher Vectors. For dimensionality reduction, we discussed methods based on feature projection and feature selection. We also discussed feature fusion and decision level fusion, which have been reported in research literature to be useful for improving performance of automated screening methods. We provided an introduction to various machine learning techniques which we use later in this thesis. Finally, we provided a brief introduction to statistical tests commonly used in SSP/AC and link them to specific applications in the field.

In the next chapter, we continue with our discussion on the development of automated screening methods. We move on from discussing fundamentals of these methods to a more targeted discussion on the development of automated methods to screen for mental and neuro-developmental disorders.

Chapter 3

Automated Screening for Mental and Neuro-developmental Disorders

3.1 Introduction

The Diagnostic and Statistical Manual of Mental Disorders (DSM) [18] is a reference book published by the American Psychiatric Association which offers a common language for the codification of mental disorders. More importantly, especially in the context of developing automated screening methods, the DSM manuals provide standard diagnostic criteria for various types of disorders, which include major depressive disorder (commonly known as depression), autism spectrum disorder (commonly known as autism), bipolar disorder, post-traumatic stress disorder, and schizophrenia amongst many others. Our aim is to utilise the standard diagnostic criteria for depression, bipolar disorder, and ASD to develop automated screening methods for these disorders and accordingly, provide discussions in Chapters 4, 5, and 6, respectively.

The DSM-5 manual (the latest edition) defines a mental disorder as a *'syndrome characterized by clinically significant disturbance in an individual's cognition, emotion regulation, or behaviour that reflects a dysfunction in the psychological, biological, or developmental processes underlying mental functioning'*. Meanwhile, *neuro-developmental disorders* are defined as an umbrella term for a group of conditions which start becoming apparent in the developmental phase of a child. According to the DSM-5, the neuro-developmental disorders are *'a group of conditions with onset in the developmental period. The disorders typically manifest early in development, often before the child enters grade school, and are characterized by developmental deficits that produce impairments of personal, social, academic, or occupational functioning.*

The range of developmental deficits varies from very specific limitations of learning or control of executive functions to global impairments of social skills or intelligence'. Relevant to this thesis, depression and bipolar disorder are considered mental disorders, whereas ASD is considered a neuro-developmental disorder.

Relevant to this thesis, depression and bipolar disorder are considered mental disorders, whereas autism is considered a neuro-developmental disorder. It is important to mention here that while the DSM-5 manual has organised disorders under special sections, some researchers have preferred to use more generic terms to describe these disorders. For example, Cohen et al. [199] refer to depression, schizophrenia, and bipolar disorders as 'serious mental illness'. Asgari et al. [200] refer to clinical depression and autism as 'cognitive impairments', and Hantke et al. [201] cluster individuals with various types of mental, neurological, and physical disabilities and refer to individuals affected by these disabilities as 'cognitively impaired'.

Our primary aim in this chapter is to discuss how features can be extracted from audio/visual modalities and subsequently used for screening of mental and neuro-developmental disorders. It must be mentioned here that our discussion will be somewhat biased towards screening of depression. This simply due to the fact that there are more publicly available datasets focusing on the task of depression screening than any other disorder, thereby gathering most attention from researchers in the SSP/AC community. Our secondary aim is to discuss inherent limitations of automated screening methods, we believe this is very important so that deliverables are reported with well-founded understanding.

The rest of this chapter is organised as follows: we start with a brief discussion of psychomotor changes which occur due to mental disorders such as depression and bipolar disorder. These changes can have a profound effect on the social signals of individuals affected by these disorders. Next, we discuss in detail how information conveyed by social signals from audio and visual modalities can be quantified as audio/visual features. As per the scope of the thesis, we focus on information extracted from the face region from the visual modality and speech from the audio modality. We then discuss in detail inherent limitations which exist and challenges which need to be overcome for the task of developing automated methods for screening mental and neuro-developmental disorders.

3.2 Psychomotor Changes

Psychomotor activities are skills which require coordination between the brain and the body to function effectively. Psychomotor activity spans multiple domains which include body movements and speech, and how well activities are affected by underlying mental processes and emotions [202].

A number of mental disorders have an adverse effect on psychomotor

activities, implying that monitoring of psychomotor activities can lead to recognition of these mental disorders. For example, depression and bipolar disorder can be recognised by identifying psychomotor symptoms as per the DSM-5 manual [18]. It is important to mention here that according to the DSM-5 manual, neuro-developmental disorders, in particular autism spectrum disorder, do not result in psychomotor changes.

In the context of mental disorders, there are two fundamental types of psychomotor symptoms. These include: (a) Psychomotor agitation, and (b) Psychomotor retardation.

Psychomotor agitation

Psychomotor agitation can be defined as an *inner restlessness or tension associated with increased motor movement*. From a motor perspective, it manifests as feelings of restlessness, such as undertaking repeated movements and fidgeting. Restlessness also leads to insomnia. From a cognitive perspective, psychomotor agitation can lead to emotions which involve high arousal and negative valence, for example anger and anxiety.

Psychomotor retardation

Psychomotor retardation is, in general, the slowing down of psychomotor activities. From a motor perspective it manifests as impaired speech (in terms of prosody, voice quality and articulation), sluggish body movements, and fatigue (which can lead to hypersomnia). From a cognitive perspective, psychomotor retardation can lead to impaired thinking and most prominently blunted display of affect and emotions.

3.3 Screening based on features from Visual Modality

Using the methods discussed in Chapter 2, one can craft features from social signals which are representative of psychomotor changes. These features can subsequently be used for recognition of mental and neuro-developmental disorders. We start our discussion with features from the visual modality

We organise features from the visual modality into two categories i.e. (a) craniofacial movement features, and (b) emotional expressivity features. We argue in light of research literature in the following section that Craniofacial Movement Analysis can be used to capture the motor perspective of psychomotor activities, whereas Emotional Expressivity Analysis can be used to test the hypothesis that individuals with mental and neuro-developmental disorders express emotions differently.

3.3.1 Craniofacial Movement

Alghowinem et al. [174, 175, 203, 204] made significant contributions to the task of automated screening of depression between 2013–2016. They study head motion, eye gaze and movement of facial landmarks as features for classifying between individuals with and without depression.

To study movement of facial landmarks, they compute an exhaustive set of pair-wise distances between facial landmark points, follow it up by computing velocity and acceleration contours from the distance measures. Finally, the contours are summarised using functionals. The authors hypothesise that since these features capture facial muscle activity, they can be useful to identify individuals with depression.

They study head motion by computing velocity and acceleration contours from headpose in terms of yaw, pitch, and roll, and summarise their values using functionals; similar to the approach used for facial landmarks. In order to estimate head pose, they first used the (POSIT) algorithm [205] with 2D facial landmarks and a 3D face reference shape to estimate the rotation matrix. Using the rotation matrix, they computed yaw, pitch, and roll.

Finally, to study eye gaze, they manually annotated up to 45 images of each participant in the database and build a 74 point active appearance model (AAM) [206] around the eye region. They then use distance measures to describe eye openings, and horizontal and vertical eye gaze.

Key findings from their research affirmed that individuals with depression do indeed show signs of reduced motor activity. These individuals have smaller facial muscle movement, slower head movement, and they tend to keep their head in the same position for a longer period of time. Interestingly, they also find contact avoidance to be a key discriminating factor, using both head pose and eye gaze features.

While it is appreciated that their work is the first major attempt in finding potentially useful visual features for depression recognition, there a number of caveats in their approach. The major caveat is to compute an exhaustive set of distance measures from 68 facial landmarks. Secondly, they used a few thousand t-tests to whittle down useful features without correction for multiple-hypothesis testing, arguing that they used t-tests for feature selection rather than hypothesis testing — which is not uncommon in the field of social signal processing (as discussed in Section 2.3.6). However, this argument does not alleviate the dangers of multiple-hypothesis testing. Also, they report results on the BlackDog dataset — their private dataset — the results, therefore, cannot be verified independently. Nevertheless, their work has laid down a foundation for future research in this area since it demonstrated that motor retardation does indeed manifest as reduced craniofacial movement, and it can be quantified automatically.

Dibeklioğlu et al. [207] also proposed the computation of features based on facial dynamics for depression recognition. Similar to the approach of

Alghowinem et al. [174,175,203,204], these authors compute movement of each facial landmark across the video recording as a time series. Thus for each video frame, there are 98 different time series (i.e. x and y coordinates of 49 facial landmarks). After applying a smoothening algorithm, they use PCA to reduce the dimensionality to 15 time series components at each time instant (i.e. each frame), retaining 95% of the total variance. Next, they compute velocity and acceleration contours. They also divide velocity and acceleration contours into segments of increasing and decreasing values before using functionals for feature summarisation. A similar procedure is used for the time-series for yaw, pitch, and roll which represents the head-pose. Instead of t-tests used by Alghowinem et al., these authors use Minimum Redundancy Maximum Relevance (mRMR) [106] for feature selection. Dibeklioglu et al. replicate the method of Alghowinem et al. [104] on their own dataset and report that their method provides better results in terms of classification accuracy. Although the results in Table 3 of their paper, Dibeklioglu et al. [207] suggest that performance improvement is likely dominated by the better performance of mRMR feature selection algorithm when compared to t-tests.

For submission to the AVEC 2016 Depression classification challenge (AVEC 2016 DCC) [12], Pampouchidou et al. [150] quantify muscle movement via motion history images (MHI) [208] of facial landmarks. Instead of using all available landmarks, they selected landmarks which represent eyebrows, eyes, nose tip, and mouth. Prior to computing MHI, they register all landmarks with respect to landmarks on the temple, chin, and inner and outer corners of the eyes. Later they compute the magnitude of change between frames, similar to the concept of histogram of optical flow (HOF) [209].

For head motion analysis, Pampouchidou et al. computed velocity and acceleration contours of four facial landmarks located on the contour of the face i.e. landmarks $\{2, 4, 14, 16\}$ for a 68-point reference. For measuring eye blinking, they compute area occupied by facial landmarks around the eyes and use an empirically determined method to measure eye blinks. Finally, they use a number of functionals to summarise velocity and acceleration contours.

Apart from facial landmarks, Pampouchidou et al. also computed functionals from facial action units (AUs) [40] and gaze features provided as part of the dataset. However, they report that both AUs and gaze functions proved detrimental to the classification performance, although MHIs computed from facial landmarks were useful. They report that visual features provide a mean F1 score of 0.58 with LOOCV on the combination of training and development partitions, 0.70 on the development partition only, and 0.47 on the test partition.

While it is appreciated that Pampouchidou et al. [150] seek to develop innovative methods to measure dynamics from visual features, however, computing features from landmarks which are towards the edge of the face contour, makes their method highly susceptible to failures of the facial landmark tracker. It may make more sense to use facial landmarks closer to the centre of the face

with the understanding that they will have minimal effect from poor tracking. For example, head motion can be gauged by using landmarks around the nose instead of contours of the face. Furthermore, their approach focuses on maximising classification accuracy and failed to provide intuition on how well their features relate to depression severity.

Yang et al. [149] was the only submission to AVEC 2016 DCC to beat the challenge baseline on the test partition. They first perform registration of landmarks using the mean shape computed from 51 stable landmarks from the training, development and test partitions. They proceed to compute distance and angle measures for eye, eyebrow, and mouth regions. PCA is applied to reduce dimensionality of the feature vector while retaining 99.90% of variance. The average value of each feature over the entire recording is taken as a global feature. They also experimented with using AUs. While their proposed features demonstrate promising results on the development partition, the authors report that their challenge winning submission was actually based on manually crafted decision trees from interview transcripts.

In their submission to the AVEC 2016 DCC, Huang et al. [155] compute multi-resolution, multi-lag auto- and cross-correlation between sequences of facial landmarks (both 2D and 3D), eye gaze, head pose, and AUs, using an approach originally proposed by Williamson et al. [210]. They achieve a mean F1 score of 0.73 on the development partition using these features, however, their best results on the test partition are from processing interview transcripts.

Nasir et al. [152] also process facial landmarks and use them as features to quantify facial muscle movement. Apart from computing distances between facial landmark points, they also compute area between certain facial landmarks. We argue that the most important deliverable from their work is the use of multi-resolution approach to feature description, where they argue that symptoms of depression may not manifest at small time intervals over which audio-visual features are typically computed. Interestingly, they claim that 'polynomial parameterisation of facial landmark features achieved the best performance among all systems', but this is not true as suggested by Table 4 in their paper. It is clear that geometric features which are based on distance and angle measures achieve better performance. Nevertheless, using a combination of visual features, they achieve a mean F1 score of 0.84 on the development partition and 0.55 on the test partition, both of which are promising.

Williamson et al. [130] computed correlation structure features of various facial action units in line with their previous works [128, 210]. They achieve a mean F1 score of 0.53 on the development partition using these features, along with a Pearson correlation value of 0.44.

Lucas et al. [177] utilised a subset of videos from the Distress Assessment Interview Corpus (DAIC) [211], they report very high correlations between frowning, smiling and eye gaze with respect to the Patient Health Questionnaire PHQ [21] scores. The caveat, however, is that their dataset contains data from only 6 participants, which compromises the validity of their results.

Nevertheless, their work provides us the motivation to undertake a cross-corpus analysis of features related to craniofacial movement.

Stratou et al. [186] also undertake analysis of a subset of the DAIC corpus, which contains individuals with depression, Post-Traumatic Stress Disorder (PTSD) or both depression and PTSD. Their results show that as the severity of depression and PTSD increases, individuals with these mental disorders show reduced facial activity in terms of AUs, head movement and eye gaze variations. It must be mentioned here, however, that specialist software was used to compute features from the face region, while other publications mentioned so far typically used non-commercial tools to compute features.

3.3.2 Emotional Expressivity Analysis

Given that individuals with mental and neuro-developmental disorders typically display altered affect, one can leverage emotional expressivity analysis to potentially identify them amongst healthy individuals.

Emotional expressivity analysis using facial expressions is based on Paul Ekman's theory for basic emotions [7]. According to him, while human beings can exhibit a very large number of emotions depending on culture, age, gender etc., there still exist a set of six 'basic' emotions which are manifest irrespective of these confounding factors. The six basic emotions are: (1) Anger, (2) Disgust, (3) Fear, (4) Happiness, (5) Sadness, and (6) Surprise. Ekman also defined these emotions in terms of specific facial muscle movements, as defined by his facial action coding system (FACS) [1, 2].

Scherer et al. [126] study anger, disgust, contempt, fear, joy, surprise, sadness, and neutral emotions computed from Computer Expression Recognition Toolbox (CERT) [84]. These test how emotional variability and emotional neutrality changes with respect to the depression severity of individuals. They report a Pearson correlation value of 0.198 for emotional neutrality with a Hedge's g effect size of 0.840. Meanwhile, emotional expressivity has a Pearson correlation value of -0.054 for emotional neutrality with a Hedge's g effect size of -0.757, thus showing that as depression severity increases, flat affect becomes dominant. In another publication [173], these authors show that anxiety, depression, distress, and PTSD, all lead to a reduction of smile intensity.

Stratou et al. [186] provide an analysis of emotions in terms of AUs rather than Ekman's basic emotions. They report that the emotional expressivity is highly dependent on the gender of the individual. They found that men have an increased activation of AU4 compared to women — which means that men tend to smile more under depression compared to women. They observe a similar trend for disgust and trend for both individuals with depression and PTSD.

Lucas et al. [177] report that individuals with depression and PTSD show facial expressions of hostility, grief and diminished signs of joy. While their observations are in agreement with research literature from psychology i.e.

psychomotor retardation, it needs to be mentioned that they experimented with data from only six individuals.

Vijay et al. [74] used AU intensity features extracted from the OpenFace toolbox for a study on facial expressions on various forms of psychosis, including depression. They report that individuals with depression show eye widening (actually larger intensity of eyelid raiser, AU5), and smaller eye openings. They also report the standard deviation of brow lowerer (AU4) intensity to be discriminatory between depressed and non-depressed individuals. They find that subjects who had a higher Brief Psychiatric Rating Scale scores (BPRS) [212], also raise their eyebrows more. While 'more' is a vague description, we do find from our own analysis on the AVEC 2016 dataset that dynamic range of eyebrow movement's velocity contour is positively correlated with PHQ-8 score for depression. They also report the following features to be useful on certain specific tasks based on the Positive And Negative Syndrome Scale (PANSS) [213]: standard deviation of AU5 intensity for focusing/thinking questions on PANSS-G6 depression scale, mean AU2 intensity for self-confidence question on PANSS-P1 delusions scale and mean value of AU12 intensity for self-confidence PANSS negative cumulative scale. A similar approach was also used by Laksana et al. [180] in their investigation on the facial behaviour indications of those with suicidal tendencies. In fact, these papers are part of the recent shift from brute force approach to maximising classification or regression accuracy and moving towards interpretable features.

While not for depression, Wortwein et al. [190] undertake analysis of facial expressions of individuals with anxiety for public speaking tasks. They used the FACET toolbox [214] to compute a number of emotions, and report on the Pearson correlation (ρ) of these emotions with respect to anxiety. They report that individuals with anxiety also display sadness with a correlation value of $r = 0.21$, while fear correlates with anxiety with $r = 0.41$.

3.4 Screening based on features from Audio Modality

While discussing the importance of exploring potential biomarkers for automated depression screening, Horwitz et al. [215] propose linking up perceptual qualities of speech from depressed individuals to the source-system (a.k.a. source-filter) model of the speech production system. Their methodology is also supported by an earlier work by DeBodt et al. [216], where they report that speech impairments can be a result of malfunctioning of one or more subsystems of speech production i.e. articulatory, resonatory, phonatory, and respiratory. We find motivation from the work of Horwitz et al., and follow their methodology to search for acoustic biomarkers for screening of mental and neuro-developmental disorders since the method provides a systematic and robust means to identify biomarkers.

According to the source-system model of speech production [217], two fundamental processes are involved in speech production. These include the generation of sound sources at the glottis or along vocal tract followed by filtering of these sources by the vocal tract. This theory suggests that speech analysis can be undertaken by means of acoustic features which represent the voice source(s) and the vocal tract.

However, in order to explore the effects of depression on speech more elaborately, Horwitz et al. add, explicitly, the category of prosody analysis in their work since it correlates with one of the most common perceptual cues of depression i.e. monotony of speech. The analysis of voice source can enable the study of the effects of mental disorders on voice quality [47, 218], while the analysis of vocal tract may identify the effects of psychomotor changes (agitation and/or retardation) [218, 219].

We organise our study on automated methods for screening mental and neuro-developmental disorders from speech into four parts:

- Prosody Analysis
- Voice Quality Analysis
- Spectral Modelling
- Formant Space Analysis

3.4.1 Prosody Analysis

Speech prosody is a generic term that describes a number of acoustic features which quantify the melody accompanying speech. These features can include but are not limited to pitch, loudness, word rate, and pauses [220].

It is a well-established feature of speech which varies in sympathy with the underlying emotions and mood of an individual [156, 221–223]. Thus, prosody is an important aspect of speech production through which one can communicate affect. Prosodic analysis has previously been used to identify individuals with depression [34,162,163,207,224,225], bipolar disorder [226,227], schizophrenia [228] and ASD [46, 229, 230].

Pitch

Pitch represents the frequency of vibrations of the vocal cords during speech production [231]. It is an important part of speech communication since it can be modulated to convey a variety of meanings, for example, emotions of happiness, sadness, or surprise and to mark an utterance as a question or an answer. Fundamental frequency is typically used as a correlate for pitch [28].

Loudness

Loudness is the psychological counterpart to sound pressure level. Sound pressure level is a physical quantity, but loudness is a psychoacoustic quantity. The former has to do with how a microphone perceives sound, the latter how a human perceives sound. It measures the sound strength. It is important for making speech intelligible and for expressing emotions. [28].

3.4.2 Voice Quality Analysis

Voice quality features capture information pertaining to the voice source, at the glottis and vocal fold level [232]. Analysis of the voice quality provides a means of gaining insight into the physiology of the glottal source.

Interestingly, there is no consensus on the number of categories for classifying voice qualities. For example, according to Gobl et al. [233], voice quality can be classified as harsh, tense, modal, breathy, whispery, creaky, or lax-creaky. We, however, follow the seminal work of Scherer et al. [126] where they use voice quality features for automated depression screening. They measure on a voice quality scale between breathy and tense, with modal voice being at the centre. The characteristics of these types of voice qualities, based on [199, 233, 234], can be summarised below.

Modal Voice: Modal voice represents normal speech. There is moderate laryngeal tension, vocal fold vibrations are efficient and the glottal flow waveform has moderate pulse.

Breathy Voice: In breathy voice, there is minimal laryngeal tension. Vocal folds, therefore, do not come fully together resulting in audible aspiration noise. The idealised glottal flow resembles a sinusoid.

Tense Voice: A tense voice results due to a higher degree of tension in the entire vocal system. Vocal fold vibrations are irregular since they are compressed tightly. This type of speech is characterised by a narrow glottal pulse with long closed phase.

The hypothesis follows that psychomotor symptoms affecting voice sources can be quantified via voice quality features. These features have previously been used to screen individuals for depression [111, 125, 126, 177, 235, 236], Psychosis [196], Suicidal tendencies [74, 237], Melancholia (a type of depression) [238], and Bipolar disorder [239].

In this thesis, we use a set of voice quality features which can be computed as part of the open-source COVAREP toolbox [62]. The COVAREP toolbox provides state-of-art algorithms for quantification of voice quality, and has been used to compute baseline features for 2016 and 2017 editions of the AVEC depression recognition sub-challenges. These features include normalised amplitude quotient (NAQ), quasi-open quotient (QOQ), parabolic spectral

parameter (PSP), Maxima Dispersion Quotient (MDQ), Harmonic-to-noise ratio (HNR), Corrected difference of first two harmonic amplitudes (H1H2), Cepstral Peak Prominence (CPP), and Peak Slope (PS). We note from the works of [199, 234, 240, 241] that all these features, with the exception of CPP have a monotonic decreasing value as voice changes from breathy to tense. CPP feature values increase as voice changes from breathy to tense.

Voice quality features can be of two fundamental types. The first type are based on estimates of glottal airflow. The second type of voice quality features are measured from speech spectra. While the features in the latter category do not directly measure glottal activity, these are known to with correlate with it [47]. This class of features includes:

***NAQ*:** The NAQ quantifies the glottal closing phase [242], and shows a monotonic decreasing trend when speech changes from breathy to creaky.

***QOQ*:** The quasi-open quotient describes the relative open time of the glottis [47]. It is measured by detecting the duration during which the glottal flow is 50% above the minimum flow. This is then normalized by the local glottal period.

***HNR*:** Harmonics to Noise Ratio quantifies the amplitude differences between the tonal peaks and the mean noise level in speech. Noise, for example, is introduced during the speech production process when vocal folds do not close completely during the closed phase [47].

***H1H2*:** It describes the amplitude of the fundamental frequency relative to that of the second harmonic. Positively correlated with breathiness in speech. The difference between the first two harmonic amplitudes of the speech spectrum (H1H2) has been shown to correlate with the open quotient of the glottal waveform [47].

***CPP*:** Cepstral Peak Prominence is the ratio between the amplitude of the cepstral peak and the overall signal amplitude [243]. It measures the degree of periodicity of speech. For example, an increase in turbulent noise leads to a decrease in the CPP value. Hence, CPP values are lower for breathy speech compared with non-breathy speech. negatively correlated with breathiness in speech [244].

***PSP*:** The Parabolic Shape Parameter quantifies the spectral decay of glottal flow [240] by fitting a parabolic curve to its signal. It has been should to be related to voice quality [125].

***PeakSlope*:** The PS involves fitting a regression line to the maximum amplitudes of speech spectra divided into octave bands. It can be used to discriminate between breathy and tense voice [241].

***MDQ*:** It describes how the spectral decay of an obtained glottal flow behaves with respect to a theoretical limit corresponding to maximal spectral decay [240].

***Jitter*:** It quantifies the irregularity in pitch. Jitter is a measure of frequency instability.

***Shimmer*:** It quantifies the irregularity in loudness. Mean absolute difference between the amplitudes of consecutive periods, divided by the mean amplitude.; shimmer is a measure of amplitude instability

It is important to mention here that in order to classify voice between breathy and tense based on a feature value, one needs information about the feature value for modal voice since feature values are known to be subject dependent. For example, consider the work of Alku et al. [240] on the use of PSP features for voice quality. While it is clear that the feature value is monotonically decreasing as voice changes from breathy to tense, the magnitude of this feature is indeed subject dependent. Thus, without a reference for the value of a particular voice feature for modal voice, any discrimination of voice into breathy or tense, is at best data driven and potentially prone to errors.

3.4.3 Spectral Modelling of Vocal Apparatus

Psychomotor changes affect muscle movements [18], these changes affect the speech production system by either increasing or decreasing vocal tract muscle tension, depending on whether retardation or agitation manifests in the individual [219, 245]. This in turn modulates spectral characteristics of voice source signals in terms of both, their magnitude and phase. To summarise, it is possible to study the effects of mental disorders on the speech production system by modelling the spectral characteristics of speech signal.

There are two fundamental approaches to spectral modelling, which we discuss briefly in subsequent sections after brief overview of commonly used spectral features of speech.

Spectral Features

Spectral features characterise speech spectra i.e. the frequency distribution of the speech signal. These features offer detailed information about the vocal tract structure and are especially useful for quantifying changes there due to muscle tension and control [246]. We provide a brief description of various spectral features as follows. These features have been used to model speech spectra.

***MFCCs*:** Mel Frequency Cepstral Coefficients (MFCCs) are arguably the most popular spectral features for speech signals. The key attribute of MFCCs

is that they represent speech spectra with respect to human perception of pitch. This is achieved by warping speech spectra with respect to the Mel scale, which relates real frequency to perceived frequency [247]. A compact *cepstral* representation of speech spectra is subsequently created using discrete cosine transformation of the Mel filtered log-energy speech spectra.

***PLPs*:** Proposed by Hermansky et al. [248], PLP features share similarities with MFCCs, even though both were developed independently [249]. While MFCCs warps spectra according to the human perception of pitch, PLP features model the perception of loudness at different frequencies [30]. This is achieved by warping speech spectra with respect to the Bark scale. Both, MFCCs and PLPs, use cepstral representation of spectra.

***Auditory Spectrum*:** Auditory spectrum (AudSpec) features are similar to PLP features in that speech spectra is warped to the Bark scale. However, unlike PLPs, AudSpec features do not use cepstral representation, instead contain summation of energy within critical bands.

***Spectral Flux*:** Spectral flux features quantify variation in the power spectrum. This is achieved by comparing the difference in power spectrum between successive short-time frames.

***Spectral Entropy*:** As the name suggestions, spectral entropy features quantify the amount of information conveyed by the power spectrum of the speech signal. Spectral entropy features represent the peakiness of speech spectra [250].

***Spectral Variance*:** Spectral variance is the second moment of statistics computed for speech spectrum, and provides information about the shape of the spectrum. It is particularly useful in differentiating between tone-like and noise-like sounds.

***Spectral Skewness*:** Spectral Skewness is the third moment of statistics computed for speech spectra, and provides quantification of spectral tilt.

***Spectral Kurtosis*:** Spectral kurtosis is the fourth moment of statistics computed for speech spectra, and provides information about the peakiness of the power distribution. A large value for kurtosis means that peaks are usually well defined, whereas a small value means that power distribution is relatively flat.

***Spectral Slope*:** This feature provides the slope parameter of a linear regression line between two cut-off frequencies of the power spectra. openSmile [30] provides spectral slope values for two ranges of frequencies i.e. (a) 0–500 Hz and (b) 500–1500 Hz.

***Spectral Energy Proportions*:** Spectral energy proportions (SEP) features quantify energy contained within certain frequency bands relative to the total energy in a short-time power spectrum. openSmile [30] provides SEP features for two bands i.e. 0–500 Hz and 0–1000 Hz.

***Spectral Roll Off Point*:** This feature provides information about the shape of speech spectrum. It determines the frequency below which a predefined percentage of the total spectral power is concentrated. openSmile [30] computes these features for 25%, 50%, 75% and 90%.

***Psychoacoustic Sharpness*:** This feature quantifies the amount of high frequency content in speech, such that perceptually, the speech sounds sharp.

***Harmonic Differences*:** These features are computed from the magnitude of harmonics of the fundamental frequencies. They represent the difference between the amplitude of a particular harmonic and the amplitude of the peak fundamental frequency. openSmile [30] provides two harmonic difference features i.e. H1–H2, which is the difference between the first and second harmonic, and H1–A3, which is difference between the first harmonic and the third harmonic.

It is important to mention here that H1–H2 has been proved to be associated with voice quality (see [47] and references therein), i.e. a breathy voice has larger values of H1–H2 compared to a tense voice.

***Alpha Ratio*:** This feature provides the ratio between the energy of speech signal between 50–1000 Hz and 1–5 kHz. In essence, it provides a measure of spectral slop. It is loosely correlated with voice quality [251].

***Hammarberg Index*:** This feature provides ratio between the peak energy in the 0–2 kHz region and the peak energy in the 2-–5 kHz region. Thus, similar to the alpha ratio, Hammarberg index also provides information about spectral slope and voice quality [252].

Spectral Harmonicity: As the name suggests, spectral harmonicity quantifies the extent of harmonics in the signal. It can be used to differentiate between, for example, voiced and unvoiced speech [253].

Spectral Modelling Methods

There are two fundamental approaches to spectral modelling. The first approach is based on using the process flow for legacy (i.e. non-deep learning) automated speaker recognition systems [95]. While the aim in speaker recognition systems is to link speech recordings to a particular individual, the objective of spectral modelling in our work is to build a model which represents speech characteristics of individuals with mental and neuro-developmental disorders.

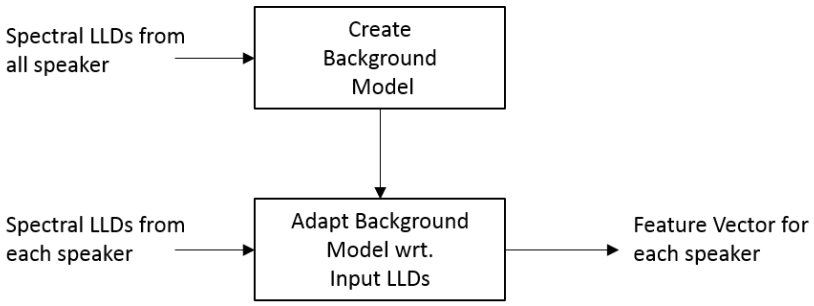


Figure 3.1: Generic Block diagram for Multi-dimensional Spectral Modelling

The generic process flow diagram for spectral modelling is illustrated in Figure 3.1. In the first stage, signal processing algorithms are used to extract spectral features from speech waveforms. As the name implies, these features provide detailed information about speech spectra and typically span multiple dimensions. The second stage involves building a generative model to learn the *feature space* of those features. The generative model, often called the background model, is trained on spectral features from speech recordings of all speakers. In the final stage, a discriminative model is built for speech recording of each speaker by quantifying the difference between spectral features from their recording and the background model. It is common to use either some or all of the parameters of the discriminative model as features for the classification or regression.

There are a number of ways to implement the generic spectral modelling process flow, illustrated in Figure 3.1. Cummins et al. [254] and Lopez et al. [255] build background models using GMM, and follow it up with GMM-UBM approach [95] to identify individuals with depression from speech. Meanwhile, Senoussaoui et al. [154], Lopez et al. [256] and Nasir et al. [152] use the i-vector approach [257] for building models for depression screening. Finally, in [34,37], we introduce Fisher Vector encoding [98] of spectral features for the task of automated depression screening. All these approaches have demonstrated success for the task of depression recognition.

There are two major advantages of this spectral modelling approach. First, it uses features which provide detailed information about speech spectra, typically in the form of MFCCs, PLP, or AudSpec features. It is also possible to concatenate any of the other features discussed earlier to create an even more detailed spectral representation in order to improve quality of data. The second major advantage is the use of the background model approach to learn the acoustic feature space — an approach which was the state of art for speaker recognition until deep learning became popular. This enables the discriminative model to already have an understanding of the acoustic feature space (including the background noise), and speaker specific models can be created without the need of large amount of data [95]. There are some inherent

flaws with this approach as well. For example, these features do not carry an intuitive meaning, therefore, while one can use multi-dimensional modelling to identify individuals with mental and neuro-developmental disorders, one cannot ascertain which characteristics of speech make them different from healthy individuals.

Speech spectra can also be modelled by functionals based feature summarisation [162]. In this approach, functionals of descriptive statistics are computed of each LLD separately and the resultant functionals are concatenated together to craft a multi-dimensional feature vector.

While trivial compared to the first approach, modelling speech spectra using functionals is interpretable [162] and is therefore more useful for experiments on datasets which are not publicly available [58]. For example, Solomon et al. [28], report that comparing mean value of MFCC coefficients 0 versus 1 i.e. MFCCs 0 vs MFCCS 1 can be used to identify individuals with depression — even if they attempt to hide their symptoms. Furthermore, Furorali et al. [46] conduct a systematic review of speech as a potential biomarker of individuals with ASD. It is clear from their review that functionals based spectral modelling is indeed useful for identifying speech of individuals with ASD.

This method has also proved successful for brute-force training on datasets. For example, Alghowinem et al. demonstrate this approach in [104] where they use t-test based feature selection on functionals to identify features which prove useful for the task of depression screening. Furthermore, authors of [28,162,224] use this approach to report on speech of individuals with depression, Wortwein et al. [196] report for psychotic people, and Chen et al. [145, 167] use this approach to report on speech characteristics of individuals with ASD.

While spectral modelling is a promising method for identifying individuals with cognitive impairments, there are some caveats which need to be acknowledged. This relates with the fact that spectral features are representative of all the information contained in speech, including both linguistic and paralinguistic — some of which may be correlated with cognitive impairment, and some of it not. It is possible that any of the many confounding factors overrides any paralinguistic information which has discriminative power [258].

3.4.4 Formant Space Analysis

An alternate method for studying the effects of mental and neuro-developmental disorders on the vocal tract is to model the formant space. Formants are the resonant frequencies of the vocal tract, therefore formant space analysis provides a direct insight into changes at vocal tract level of the production system. Formant space analysis has widely been used by the computational linguistics community, especially for understanding articulation [217, 259].

Based on the hypothesis that psychomotor symptoms affect the vocal tract functions, it follows that individuals with mental disorders will have, depending on the severity of their illness, impaired articulation. This is in part supported

by formant space analysis and it has accordingly been proposed as a robust identifier for a number of mental disorders, neuro-developmental disorders, and cognitive impairments such as depression [128, 162, 260], autism [46] (with references therein), Schizophrenia [199, 261], Alzheimer's disease [185, 262], and bipolar disorder [184, 199].

There are practical constraints, however, which hinder the application of formant space analysis to its full potential. For example, it is highly dependent on the lexical content of speech and requires phonetic segmentation as a pre-processing step. This is highlighted in the work of [128], who report that their method works best for scripted speech. In fact, they only used scripted speech recordings of the AVEC 2014 depression recognition sub-challenge where they achieved the highest accuracy. For the 2016 edition of the challenge, where they were not able achieve phonetic segmentation, the results of their formant analysis method were not particularly good [130].

3.5 Limitations and Challenges

Automated screening for mental and neuro-developmental disorders is still very much a work in progress. There are a number of limitations which need to be acknowledged and challenges which need to be overcome in order to develop such systems. In subsequent sections, we discuss some of the outstanding issues relevant to the field. These are:

- Availability of datasets
- Size of datasets
- Social Signal Modalities
- Ground truth labels
- Confounding Factors
- Dataset Creation and Design
- Choice of Software

3.5.1 Availability of Datasets

An inherent limitation in the field of SSP/AC is the lack of publicly available datasets, which inhibits research progress. We find that this issue is even worse for the task of developing automated screening methods for mental and neuro-developmental disorders.

The availability of datasets is limited to due ethical restrictions which while permitting some researchers to collect and use data, do not permit them to share data with other researchers, mostly due to privacy concerns. For

example, as part of our work on the development of automated screening methods for ASD, we sought access to datasets used by other researchers in the field. Upon contact, Bone et al. [195] and Baird et al. [144] informed us that their dataset cannot be shared with researchers outside of their group due to ethical restrictions. Meanwhile, organisers of Interspeech ComPare 2013 Autism sub-challenge [64] and Motlagh et al. [263] did not return our emails. We would like to add here that ethical restrictions do not permit us to share the dataset used by us in our work with researchers outside our group as well.

On the other hand, we note that datasets which are publicly available have drawn interest from the community, especially when it comes to the task of developing automated screening methods. For example, consider the case of datasets provided as part of the annual Audio/Visual Emotion Challenge (AVEC), which is organised as a workshop co-located with the ACM Multimedia Conference. One of the highlights of AVEC has been its challenges on development of automated screening methods for depression. As of 20th of August 2018, the introductory paper for AVEC 2013 challenge has been cited 189 times, AVEC 2014 has been cited 153 times, AVEC 2015 has been cited 80 times, AVEC 2016 has been cited 125 times, and AVEC 2017 has been cited 19 times so far. An interesting point to raise here is that AVEC 2015 did not have a challenge dedicated to depression, and when compared with other challenges, its introductory paper has the smallest number of citations.

The most recent edition of the challenge i.e. AVEC 2018 provides for the first time a publicly available dataset for the task of screening for bipolar disorder. We expect this dataset to generate interest for the development of automated screening methods for bipolar disorder, just as previous AVEC challenges helped generate interest for research on depression.

3.5.2 Size of Datasets

Most datasets available for training automated screening methods contain a relatively small amount of training data in terms of both, subjects and samples. In order to convince the reader, we provide a summary of number of subjects in each dataset from recent publications advocating automated screening methods in Table 3.1.

Here, it is shown that most datasets contain only a small number of subjects, with further categories based on ground truth labels and training, development, and testing partitions. This makes it difficult to ascertain the efficacy of automated screening methods simply because the trained models are vulnerable to the effects of overfitting.

For example, consider the case of the recently concluded Audio Visual Emotion Challenge (AVEC 2018) [33], where we participated in the Bipolar Disorder sub-challenge. Based on our work on the development of automated screening methods for bipolar disorder (see Chapter 5), we observed a trend where there is a significant decrease between the accuracy achieved on the

Table 3.1: Typical size of datasets available for developing automated methods for screening of mental and neuro-developmental disorders

Reference	**Year**	**Theme**	**Subjects and Labels**
Bhatia et al. [238]	2017	Melancholia (depression)	39, with 13 control, 13 melancholic depression, and 13 non-melancholic depression
Venek et al. [237]	2017	Suicidal risk assessment	60, with 30 suicidal and 30 non-suicidal
Wortwein et al. [196]	2017	Psychosis	20, all with psychosis
Coco et al. [69]	2017	Autism	10, with 5 typically developing and 5 with autism
Pokorny et al. [145]	2017	Autism	20, with 10 typically developing and 10 with autism
Vijay et al. [74]	2016	Psychosis	18, all with psychosis
Tron et al. [73]	2016	Schizophrenia	67, with 33 control and 34 with schizophrenia
Valstar et al. [12]	2016	Depression	142, continuous depression scale
Alghowinem et al. [175]	2016	Depression	60, with 30 control and 30 depressed
Wolters et al. [197]	2015	Depressed mood	14, all depressed
Hussenbocus et al. [225]	2015	Depression	139, with 71 non-depressed and 68 depressed
Solomon et al. [28]	2015	Depression	17, with 8 non-depressed and 9 depressed
Valstar et al. [11]	2014	Depression	84, continuous depression scale
Stratou et al. [186]	2013	Depression and PTSD	53, with overlapping illnesses with 22 PTSD positive, 31 PTSD negative, 17 depression positive, and 36 depression negative
Scherer et al. [125]	2013	Depression and PTSD	30, with 10 control, 10 suicidal, and 10 depressed

development partition compared to the test partition. However, given the relatively small size of the dataset, one does expect such observations.

3.5.3 Social Signal Modalities

For the development of systems which can complement trained clinicians, automated systems need to have access to same information that is available to a clinician. For example, clinicians can both, watch and hear, the patient during the interview process. Thus, clinicians have access to information about the patient's facial expressions, head motion, body posture, and speech. In most cases clinicians will also have access to the patient's medical history, including any medications they take.

However, datasets currently available for developing screening methods simply do not have access to the same number of modalities. For example, datasets available as part of 2013, 2014, 2016, and 2017 editions of the AVEC depression recognition challenges only contain information about the patient's face and their speech [10–13]. Naturally, with limited modalities, one cannot expect automated systems to reach performance of clinicians. In the worst case scenario automated systems may even start making false conjectures based on missing social signals.

3.5.4 Ground Truth Labels

While automated screening systems can indeed bring objectivity to the screening process, these systems are still trained on data which has inherently subjective labels. We find this common in many applications of social signal processing, especially those in the domain of automated screening of mental and neuro-developmental disorders. For example, it is common to use self-assessment forms to provide labels for automated screening of depression severity [10–13, 28], bipolar disorder [33, 39], and ASD [46].

Let us visit the case of ground truth labels used for the task of automated screening of depression in more detail. In most cases, labels for depression severity scores are based on self-assessment forms, which essentially rely on individuals to honestly report on the questionnaires. This may not always be true. In fact, for the AVEC 2017 challenge, we note that certain participants have a Patient Health Questionnaire (PHQ) [21] score of zero – which may show that they have excellent mental health – but in the interview transcripts these participants go on to discuss their battles with depression and post-traumatic stress disorder in the past. To hone in on the point of potentially noisy labels in the dataset, consider the case of participant with ID 464, who has a PHQ score of 0. From the interview transcript, one finds this individual saying, '*I know how it's like to be depressed . . . how does depression feel like . . . like a bird in a cage . . . a fish who can't swim in water . . . a bird without wings . . . like you're limited*'. When pressed by Ellie (the virtual avatar) the participant finally concedes, '*I could say today, you know, earlier when I was just by myself I felt a little depressed*', even though the PHQ score for this individual is 0.

The use of ground truth labels based on self-assessment forms mean that

labels are potentially noisy, which harms the machine learning process. However, as noted by Solomon et al. [28], despite the inherent flaws of self-assessment forms, they provide at least a reasonable and quantifiable standard to measure depression against. Furthermore, it needs to be appreciated that at this stage automated methods do not have objective labels. In the case of depression screening, computerised algorithms, at best, predict self-assessment forms. This limitation, however, comes directly from the field of psychology, since mental and neuro-developmental disorders are at best diagnosed based on behavioural symptoms [28, 33, 264].

Therefore, any system trained on the data currently available publicly is more accurately described as predictor of depression scores rather than one detecting depression. It is essential to distinguish between these descriptions in order to avoid misrepresentation. This is because predicting scores of a self-report questionnaire does not strictly mean detecting depression and vice versa. While it would certainly help if datasets also include reports from trained clinicians as ground truth labels along with self-report forms, however, to the best of our knowledge no publicly available dataset for the task of depression screening includes this information.

3.5.5 Confounding Factors

The limitations due to relatively small size of datasets is further exacerbated by many confounding factors which can affect social signals in addition to possible impairments due to mental and neuro-developmental disorders. This in turn affects the accuracy of automated screening methods.

For example, McDuff et al. [265] conducted a large scale study on cultural differences in observed facial behaviour based on 740,984 subjects from 12 countries around the world in which they report that culture and gender independently play a major role in the characteristics of facial affect. Similarly, acoustics of speech can be affected by various confounding factors such as climate, language, ethnicity, the lexical content, age, gender, and medication status amongst others, as per discussion by Schuller et al. [266].

Sagha et al. [258] investigated the effect of age, gender, and OCEAN personality traits (Openness to experience, Conscientiousness, Extroversion, Agreeableness, and Neuroticism) on the performance of automated emotion recognition from speech. They found that age is the biggest confounding factor, closely followed by gender. Amongst the OCEAN personality traits, conscientiousness and neuroticism had the most effect. An earlier work by Kring et al. [267] also found gender to be a significant confounding factor.

Gross et al. [268] reported the difference between induced and naturally occurring emotions. They highlight that while emotions can be induced, by showing affective pictures or music, an individual may not have the same response in terms of changes in facial muscles or vocal tract as they would with natural emotions. We find deliverables from their work important since

several popular datasets which are used for developing automated screening methods use induced emotions, for example the dataset used for the ComParE Self-Assessed Affect sub-challenge [148].

We find that existence of confounding factors has been acknowledged by a few members in the research community for the task of developing automated screening methods [71], however, the scale of effect is disputed. For example, while [104, 111, 149] report that gender is a significant confounding factor for automated screening of depression from speech modality, research from Cummins et al. [162] suggests that gender differences exist only when detailed (i.e. high-dimensional) spectral features are used, implying that gender does not act as a confounding factor when single-dimensional features are used. Honig et al. [127] disagree, and report that even single-dimensional features are affected by gender.

Similarly, Alghowinem et al. [109] and Williamson et al. [128] make differing conclusions about the usefulness of spontaneous speech or scripted speech for automated screening for depression, with Alghowinem et al. reporting in favour of spontaneous speech while Williamson et al. reported in favour of scripted speech. It is, however, clear from the work of Price et al. [269] that different brain regions are activated by spontaneous and scripted speech, implying that different sets of speech features may be useful for these two types of speech.

It is important to mention here that confounding factors can also exist due to differences in recording environments. For example, the dataset used for the ComParE 2013 Autism sub-challenge [64] contained speech recordings from typically developing children and atypically developing children. One of the objectives for the challenge was to develop machine learning models to differentiate between the speech of the two groups of children. As it turned out that atypically developing children were recorded in hospital environment, whereas the children from the typically developing group were recorded in a variety of different environments. As part of their solution to the challenge, Bone et al. [270] demonstrated that it is possible to differentiate between the two groups of children by explicitly focussing on the acoustics of background in the speech spectra. Thus, through their experimentation they highlighted the influence of recording environment acoustics as a major confounding factor.

Karasz et al. [271] investigated the role played by cultural differences in screening of depression. Here, they argue that depression cannot be studied without first taking into account both, the social context and environment, since factors such as nationality, ethnicity, and socio-economic status influence the prevalence and the presentation of depression. However, to the best of our knowledge, no publicly available dataset which can be used to train a model to recognise depression, provides such detailed level of meta-data.

3.5.6 Dataset Creation and Design

While dataset creation and design is a difficult and complicated task, any flaws in the design process eventually make the development of automated screening methods more challenging. As discussed previously, there exist inherent limitations due to imprecise ground truth labels, a myriad of confounding factors, and the relatively small size of datasets. In addition to these, we believe that the following aspects of the design process need to be considered as well.

The first point we raise, addresses the methodology through which raw (unedited) audio recordings are segmented to create audio files which form the dataset. Speech segmentation is one of the first steps in dataset creation where the objective is to separate parts of the audio recording which contain speech belonging to subjects and speech belonging to non-subjects. We find that this is typically achieved through manual annotation and subsequent segmentation of these recordings. For example, in the recently concluded ComParE 2018 Atypical-Affect sub-challenge [148], the dataset contained audio files which were segmented from raw audio recordings after manual annotation, details of which were not provided even after our contact with the authors of the dataset [134]. Similarly, in their work on development of automated methods for screening of ASD, Motlagh et al. [263] state that they performed manual segmentation of raw speech recordings — again, without providing further details.

We believe that a major drawback of purely manual annotation is that speech segments resulting from manual annotation and subsequent segmentation not only depend on speech activity of subjects but are also biased due to the mood, concentration, and overall subjectivity of the annotator. For example, we find that the duration of audio files for the ComParE 2018 Atypical-Affect sub-challenge range between 143 ms and 200 seconds, which is quite a large range. Furthermore, we find that in some audio files, speech from non-subjects is also not properly removed. We posit that speech segmentation can have significant impact on the deliverables of automated screening methods, and conduct experiments to investigate this as part of our work on automated screening of ASD (see Chapter 6).

On a related matter, we note that for ComParE 2017 and 2018 challenges, the time duration of some audio files is as short as 143 ms, which appears to be very short. While we do appreciate that human beings have the ability to predict human behaviour even from short duration exposure to relevant social signals [272], there is still no consensus in the SSP/AC community about how effectively machines can learn from the same. In the SSP/AC community, short duration exposure to social signals is typically referred to as *thin-slices*, and numerous experiments have been conducted to understand aspects of human behaviour using the thin-slices approach [224, 273–275]. For example, Alghowinem et al. [224] report that for the task of screening for

depression, thin-slices of speech as short as 1.4 seconds works better than taking features from entire speech recording. Similarly, Jaques et al. [275] report using thin-slices of up to 1 minute in duration for predicting bonding in conversation using speech of subjects.

We also find that various regimes are used for arranging audio/visual recordings into a dataset. For example, organisers of the AVEC challenges (editions 2013 through 2018) provide entire audio/visual recordings without segmentation i.e. they provide recorded data after truncation at the beginning and end of recording sessions. Thus, each recorded session is stored within a single file, which provides useful information for participants in terms of session number and speaker identity. When organisers have felt the need to provide information about speech segments, they explicitly provide time stamps as part of meta-data (as is the case for 2016–2018 editions of the AVEC challenges [12, 13, 39]. On the other hand, organisers of ComParE 2017 [146] and 2018 [148] performed segmentation themselves and provide segmented audio as separate files in the dataset. Moreover, we find these audio files are numbered in an apparently random manner (at least to participants of the challenge), therefore, one cannot link these audio files to either a particular subject or a recording session. Thus, for AVEC challenges, we have access to data for each subject which is longer in duration compared to the ComParE challenges.

Another aspect which plays an important role in dataset design process is how subjects are allocated between training, development, and test partitions. Ideally one expects that subjects are allocated in such a way that confounding factors such as subjects' age and gender are evenly distributed amongst the three partitions. This is in addition to the requirement of an identical distribution of labels between the partitions. One expects that machine learning models can be trained on the training partition, their hyper-parameters can be optimised using the development partition, and performance can be tested on a previously unused test partition. However, we find a lack of consistency in this regard between datasets released as part of AVEC and ComParE challenges. For example, we note that while the datasets for AVEC 2013 [10] and 2014 [11] had some subjects repeated multiple times across multiple partitions, AVEC 2016, 2017, and 2018 editions did not repeat subjects between multiple partitions. Meanwhile, datasets provided as part of ComParE 2017 and 2018 do not contain meta-data which can help ascertain whether subjects do or do not repeat between partitions, although as part of our work on ComParE 2017 Cold sub-challenge, we suspect that some subjects do at least repeat across training and development partitions.

Nevertheless, given limited availability of datasets, one often has no choice but to work with the datasets which are available, even if there are aspects which make development of automated screening methods more challenging and the deliverables imprecise.

3.5.7 Choice of Software

Kiss et al. [276] highlighted in their work that the choice of software tools can significantly affect experimental results. They found pitch features to be useful (statistically significant) for the task at hand using one software but not on another. This is most likely due to differences between implementation of signal processing algorithms, but nevertheless, it affects reproducibility of experiments if precise details of software implementation are not provided. In [46], Fusaroli et al. conduct a systematic literature review and meta-analysis as part of their investigation into the efficacy of voice as a marker for ASD. One of the key deliverables of their work is that choice of software and lack of details for it has been one of the limiting factors which has hindered research progress.

We believe that the solution to this hindrance is to either use open source software tools (i.e. standard feature-sets and standard machine learning toolkits) or for researchers to make their codes public.

3.6 The Need for Interpretability

Modelling human behaviour is an inherently complicated process [266]. As already discussed in this chapter this task is further complicated due to numerous factors, which include limited amount of data, relatively small size of datasets, imprecise ground-truth labels, limitations with respect to availability of social signal modalities, a myriad of confounding factors which affect social signals as well as subject bias introduced by dataset creation and design process.

A potential way forward is discussed by Cohen et al. [199, 277], where they emphatically argue in favour of informed research progress in terms of interpretable features and machine learning algorithms. Their main argument is that the absence of informed research progression can lead to misleading deliverables from the field of SSP/AC.

We believe so too, especially in light of the surveyed literature in this thesis and experiments conducted as part of our work on the development of automated screening methods for depression, bipolar disorder, ASD, and paralinguistic activities. It appears that other researchers also have had this realisation. There is a growing trend amongst researchers working towards the development of automated screening methods [28, 69, 73, 74, 145, 196, 237] where the authors report on the efficacy of interpretable audio-visual features rather than focussing on brute-force approaches to fit datasets.

3.7 Summary

In this chapter we continued our discussion on automated screening methods from Chapter 2, but here the discussion was more targeted towards screening of mental and neuro-developmental disorders.

We provide a summary of discussions from this chapter as follows:

- We started with a brief discussion of psychomotor changes which occur due mental disorders such as bipolar and depression. These changes can have a profound effect on social signals of affected individuals.

  We follow this up with a detailed discussion on how one can craft features from social signals from audio/visual modalities which are representative of psychomotor changes. These features can subsequently be used for recognition of mental disorders.

- We note that while the DSM-5 manual [18] has organised disorders under special sections, some researchers have preferred to use more generic terms to describe these disorders. For example, depression, schizophrenia, and bipolar disorders have their own categories, but, Cohen et al. [277] refers to these as 'serious mental illness'. Similarly, Asgari et al. [200] refer to clinical depression and autism spectrum disorder (ASD) as 'cognitive impairments'. Hantke et al. [201] cluster individuals with various types of mental, neurological, and physical disabilities and refer to individuals affected by these as 'cognitively impaired'.

- We discuss in detail current limitations of automated screening methods which need to be acknowledged and challenges which need to be overcome for further development of these methods. The challenges include limited amount of data, relatively small size of datasets, imprecise ground-truth labels, limitations with respect to availability of social signal modalities, myriad confounding factors which affect social signals, and subject bias introduced by dataset creation and design process.

Overall, in light of discussion in this chapter, we believe that development of methods for automated screening of mental and neurological disorders can greatly benefit from collaboration between researchers from SSP/AC community and clinicians.

Chapter 4

Automated Screening for Depression

4.1 Introduction

Depression is a mental illness, which according to the World Health Organisation (WHO) affects more than 300 million individuals worldwide and is the leading cause of disability [278]. At its worst, depression can trigger thoughts of suicide and is directly blamed for around 800,000 deaths every year [278]. Individuals who suffer from depression often face unemployment due to their inability to work and are susceptible to alcohol abuse [279]. Furthermore, long-term depression also increases the risk of dementia and Alzheimer's disease [280, 281].

Individuals with depression may exhibit a wide range of symptoms. In fact, Stanford University's Mood and Anxiety Disorders Laboratory [282] defines depression as a 'whole-body' illness, which affects body movements, mood, and thoughts. Depression modulates the way individuals eat and sleep, and the way one thinks about oneself and others. Sobin et al. [202] found that depressed individuals differ from non-depressed control groups in terms of gross motor activity, body movements, speech, and motor reaction time. It has been reported by Niedenthal et al. [283] and Schrijvers et al. [284] that depression adversely affects emotional state of an individual along with their physical movements, while Solomon et al. [28] report that depression is correlated with the breakdown of normal social interactions.

Automated depression screening can provide insights into the depressive state of individuals in two fundamental ways. The first is to consider depression screening as a classification task, where the objective is to determine whether or not an individual is depressed. The second approach is a regression task, where the objective is to predict the depression severity score.

In this chapter we discuss our proposed approaches for automated screening for depression from audio/visual modalities. Accordingly, we organise the rest of

this chapter as follows: we start with statements of novelty and contributions through our work. We follow this by describing traits of individuals with depression according to the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) [18]. We also discuss various depression measurement instruments which are used to gauge severity of depression. This provides the foundation for our proposed automated screening methods. Next, we describe datasets used in this thesis and work carried out by other researchers in the field for these datasets. We follow this up with detailed discussion on our proposed methods, which is supported by experimental analysis and participation in the AVEC 2017 Depression severity prediction challenge [13]. Finally, we end the chapter with a summary of key achievements of our work.

4.2 Novelty and Contributions

We summarise our contributions for the task of automated screening of depression as follows:

- We surmise that psychomotor changes due to depression lead to uniqueness in an individual's speech pattern which manifest as sudden and erratic changes in speech feature contours. To this end, we propose a novel set of temporal features, which we call *turbulence features*, to quantify fluctuations in contours of speech features. The efficacy of these features is demonstrated in terms of our solution for the AVEC 2017 depression severity prediction sub-challenge, as reported in [34].

  The reader is referred to Section 4.7.1 for details of turbulence features and to Section 4.8.2 for experimental results.

- We detail a methodology to quantify specific craniofacial movements, which we hypothesise could be indicative of psychomotor retardation and hence depression. The efficacy of these features is demonstrated by predicting test partition labels two publicly available datasets from AVEC challenges on depression severity prediction i.e. the AVEC 2014 Depression severity prediction challenge (AVEC 2014 DSC) [12] and the AVEC 2017 Depression severity prediction challenge (AVEC 2017 DSC) [13].

  The reader is referred to Section 4.7.2 for details of craniofacial movement features and to Section 4.8.1 for experimental results.

- We hypothesise that individuals with depression have unique characteristics to their speech spectra due to psychomotor changes at vocal tract level. To this end, we introduce Fisher vector encoding of spectral LLDs for quantifying abnormalities within speech spectra of individuals with depression.

Initially, we demonstrate the efficacy of our proposed approach for the AVEC 2016 Depression Classification Challenge (DCC) dataset [12], where the objective was to identify individuals with and without depression [37]. Later, we extended the idea by adding temporally-piecewise aggregation of Fisher vectors as part of our solution to the AVEC 2017 DSC [34]. We beat the challenge baseline whilst using this method.

The reader is referred to Section 4.7.3 for details of Fisher Vector features and to Section 4.8.3 for experimental results.

4.3 Depression as per the DSM-5 manual

Depression can affect an individual in many ways making it difficult to be diagnosed perfectly. There exists a need, however, for formal specification of the core symptoms of depression. The most widely accepted documentation of these symptoms exists in the Diagnostic and Statistical Manual of Mental Disorders (DSM) manual [18], which is a guidebook from the American Psychiatric Association. It is important to mention here that the DSM-5 formally defines depression as 'major depressive disorder (MDD)', however, one finds that the term 'depression' is more commonly used. Therefore, we continue to refer to MDD as depression.

According to the DSM-5 manual, an individual may be diagnosed with depression if he or she exhibits at least five symptoms from a list of nine, including at least one of the first two symptoms. These symptoms must also prevail over a duration of two weeks or more. We list the symptoms for diagnosis of depression as specified in the DSM-5 manual as follows:

1. Depressed mood most of the day or nearly every day.
2. Markedly diminished interest or pleasure in all, or almost all, activities most of the day or nearly every day.
3. Significant increase or decrease in appetite leading to unintentional weight gain or loss.
4. Insomnia or hypersomnia nearly every day.
5. Noticeable psychomotor retardation or agitation nearly every day.
6. Fatigue or loss of energy nearly every day.
7. Feelings of worthlessness or excessive or inappropriate guilt.
8. Diminished ability to think or concentrate, or indecisiveness, nearly every day.

9. Recurrent thoughts of death (not just fear of dying), recurrent suicidal ideation without a specific plan, or a suicide attempt or a specific plan for committing suicide.

The DSM-5 manual provides further guidelines to corroborate the existence of these symptoms. For example, the manual makes it clear that cues of psychomotor retardation or agitation must not merely be subjective feelings of being slowed down (in case of retardation) or restlessness (in case of agitation). These cues also need to be observable by others.

The manual acknowledges the fact that some of these symptoms may vary according to the age of the patient. For example, depressed mood may manifest in adults as feelings of sadness, emptiness and hopelessness, while for children and adolescents, it may manifest as irritable mood. Furthermore, given that some of these symptoms may be modulated by patient's history and cultural norms, the decision on whether or not an individual has depression inevitably requires the exercise of clinical judgement.

While the manual lists the most fundamental symptoms of depression, it is clear to see why it is difficult to diagnose depression perfectly: depression can have effects on two extremes, such as insomnia or hypersomnia, psychomotor retardation or psychomotor agitation, appetite loss or appetite gain. As one can expect, the absence of clear symptomisation of depression does adversely affect the process of measuring and quantifying the depressive state of an individual.

4.4 Depression Measurement Instruments

Similar to related serious mental health illnesses [199] such as schizophrenia, bipolar disorder, and post-traumatic stress disorder, there is a lack of physical means to measure depression [19, 152]. Diagnosis and assessment of symptom severity for depression is based on subjective reports from either patients themselves, their family members, or in the best case scenario, by a clinician.

Depression measurement instruments exist as multiple choice questions (MCQ) style question-answer forms. Some of these instruments are administered by a clinician as the patient is interviewed, while others instruments are self-administered forms which are filled in by the concerned individuals (aka patients) themselves. Depression measurement instruments are also referred to as *depression severity scales* or depression scales.

The most commonly used depression measurement instruments include the Hamilton Rating Scale for Depression (HRSD) [20] often abbreviated as HAM-D, the 2^{nd} edition of the Beck Depression Inventory (BDI-II) [38], the Quick Inventory of Depressive Symptomatology (QIDS) [285], and the Patient Health Questionnaire (PHQ-8) [21]. This list is not exhaustive and the reader is directed to [286] for a discussion on other depression measurement instruments.

In Table 4.1 we summarise four depression scales i.e. HAM-D, BDI-II, QIDS and PHQ-8 based on information provided in [20, 21, 38, 71, 129, 193, 207, 256, 285, 286]. Our decision to specifically select these four instruments is based on the fact that most researchers in SSP/AC have used these instruments to measure depression, as will become evident in this chapter. It therefore makes sense to have an understanding of these depression measurement instruments before we go any further.

As evident from Table 4.1, the decision to select a depression measurement instrument is not very clear and depends on a number of factors. The most important decision perhaps is to decide whether the instrument is administered by the clinician or self-administered by the patient. While clinician-administration will require time from a trained clinician, it is less susceptible to bias from the individual. The number of questions in the instrument can also be a deciding factor since it translates into time and effort required to complete the questionnaire by a clinician or the patient.

The detail in which each symptom of depression is probed in the instrument can also play an important part in the decision process, because as discussed in [287,288], some of these instruments probe more than just the core symptoms of depression mentioned in the DSM-5 manual. For example, in addition to directly querying depressive illness, the BDI-II instrument also queries feelings of being punished, pessimism, and irritability, whereas the HAM-D instrument covers anxiety, hypochondriasis, and genital symptoms. These symptoms are not, in fact, part of the DSM-5 manual for depression screening. Nevertheless, Fried et al. [287] report that both DSM and non-DSM symptoms show significant correlation for the task of depression screening.

Table 4.1: Summary of depression measurement instruments

	HAM-D	**BDI-II**	**PHQ-8**	**QIDS**
Introduction year	1960	1996	2002	2003
Instrument method	Clinician administered	Self-assessment	Self-assessment	Clinician administered (QIDS–C) or Self-assessment (QIDS–SR)
Free to use	Yes	No	Yes	Yes
Number of questions and range of score	21 questions in total but only first 17 questions are used. 8 questions with range 0–4 each, and 9 questions with range 0–2 each	21 questions with range 0–3 each	8 questions with range 0–3 each	16 questions with range 0–3 each
Time required	15–20 minutes	5–10 minutes	2–3 minutes	5–10 minutes
Total scoring range	0–7: normal, 8–13: mild, 14–18: moderate, 19–22: severe, 23–27: very severe	0–13: minimal, 14–19: mild, 20–28: moderate, 29–63: severe	0–4: not significant, 5–9: mild, 10–14: moderate, 15–19: moderately severe, 20–24 severe	0–5: not significant, 6–10: mild depression, 11–15: moderate, 16–20: severe, 21–27: very severe
Dataset(s)	Pittsburg	AVEC 2013, AVEC 2014	AVEC 2016, AVEC 2017	BlackDog

4.5 Datasets

In our work we use three datasets provided as part of challenges for automated screening for depression from the 2014, 2016, and 2017 editions of the AVEC challenges. We provide brief details of these datasets as follows:

4.5.1 AVEC 2014 DSC

The dataset for the AVEC 2014 Depression severity prediction challenge (AVEC 2014 DSC) consists of video recordings of subjects with a human computer interaction (HCI) environment. Subjects were instructed to perform two tasks as they were being recorded through a webcam and a microphone. The first task, hereby called *Northwind*, is a scripted speech task where subjects read aloud an excerpt from the fable Die Sonne und der Wind (The North Wind and the Sun). The second task is a spontaneous speech task where subjects respond to one of a number of questions such as: 'What is your favourite dish?'; 'What was your best gift, and why?'; 'Discuss a sad childhood memory'. All subjects speak German language in the dataset.

The dataset contains 150 video recordings from 84 subjects. As obvious, some of these were recorded multiple times. 18 subjects appear in three recordings, 31 in two recordings, and 34 in only one recording. The mean age of subjects is 31.5 years, with a standard deviation of 12.3 and a range between 18 to 63 years. The duration of recordings provided as part of the dataset lie between 6 and 248 seconds. In total there are 300 video recordings as Northwind/Freeform task pairs, divided as 100 recordings for training, 50 for the development, the remaining 150 recordings for the test partition.

Prior of each recording session, subjects were asked to complete the self-administered BDI-II questionnaire [289]. Scores from BDI-II are used to quantify depression severity. The objective of the AVEC 2014 DSC was for its participants to predict the BDI-II scores for training, development, and test partitions.

Participants of the sub-challenge were judged on the basis of the performance of their method in terms of the root mean square error (RMSE) between the ground-truth BDI-II scores and predictions on the test partition. For further details, we refer to the reader the baseline paper for AVEC 2014 [11].

4.5.2 AVEC 2016 DCC

The AVEC 2016 Depression classification challenge (AVEC 2016 DCC) [12] uses the Distress Analysis Interview Corpus – Wizard of Oz (DAIC-WOZ) dataset, which is part of a larger corpus i.e. the Distress Analysis Interview Corpus (DAIC) [211]. The DAIC-WOZ database consists of video recordings of clinical interviews of subjects with a virtual agent called *Ellie*. The interview structure

was designed to support the diagnosis of psychological distress conditions such as anxiety, depression, and post-traumatic stress disorder.

Due to ethical restrictions, however, organisers did not actually provide video recordings. Instead they provided speech recordings of communication between participants and Ellie along with the transcripts. The duration of these recordings range between 7–33 minutes. As baseline audio features, the organisers provide prosody, voice quality, and formant features computed from the COVAREP toolkit [62]. They also provided a set of visual features computed using the OpenFace toolkit [32]. These features provide information about the location of facial landmarks, facial action units [40], and histogram of oriented gradient (HOG) features [78].

Each subject in the dataset was required to complete a self-administered depression assessment questionnaire based on the PHQ–8 [21], which provides a measure of current state of depression. Based on the guidelines of the PHQ-8, each participant was allocated a binary ground-truth label of either *depressed* or *not-depressed*. The objective of the AVEC 2016 DCC was for the participants of the challenge to predict these ground truth labels.

The DAIC-WOZ dataset used for AVEC 2016 DCC contains multimodal data from 189 subjects in total, with 107, 35, and 47 subjects in training, development, and test partitions, respectively. Classification accuracy is measured in terms of average F1 score for prediction of subjects as either depressed or not-depressed for each partition. The F1 score is typically used as a measure of classification accuracy when large imbalance exists between class labels and is computed as the harmonic average of the precision and recall [290].

For further details, we refer the reader to the baseline paper for the AVEC 2016 DCC [12] and the supporting documents which are available within the dataset.

4.5.3 AVEC 2017 DSC

The AVEC 2017 Depression severity prediction challenge (AVEC 2017 DSC) [13] uses the same dataset i.e. the DAIC–WOZ dataset, as used for the AVEC 2016 DCC. However, the objective of the AVEC 2017 DSC was to predict depression severity in terms of the PHQ-8 scores from the self-administered depression assessment forms (i.e. a regression task) rather than the classification task of AVEC 2016 DCC. Similar to the AVEC 2014 DSC, the accuracy of submissions for AVEC 2017 DSC was measured in terms of RMSE for training, development, and test partitions.

For further details, we refer the reader to the baseline paper for the AVEC 2017 – DCC [13] and the supporting documents which are available within the dataset.

4.6 Literature Survey

Here we discuss work from other researchers for the three AVEC challenges. It must be mentioned here that research literature surveyed in this section is in addition to that already surveyed in Chapter 2 and 3.

4.6.1 AVEC 2014 DSC

Gupta et al. [223] hypothesised that head and overall facial movement carry important information about affective state and depression level. They quantify this movement using local binary patterns (LBP) [291], local binary patterns – three orthogonal planes (LBP-TOP) [81], and optical-flow-based motion vectors [209] between a pair of consecutive frames at key points computed using a corner detection algorithm [292]. They also computed pairwise Euclidean distances between each of the 66 facial landmarks and a stable point between the eyes.

For audio modality, Gupta et al. used the baseline audio features provided as part of the challenge, along with a variety of other features they had proposed in one of their prior works [293]. In total, the dimensionality of their feature vector from audio-visual modalities reached 42,092! — at which point they used brute-force feature selection based on the sequential forward search (SFS) and sequential backward search (SBS) algorithms [100]. Support vector regression (SVR) [116] was used for the task of mapping feature values to the BDI-II scores. The best result they achieved on the test partition was an RMSE = 8.99, with an RMSE = 9.68 on the development partition.

Kachele et al. [294] demonstrated that abstract meta-knowledge can be useful in prediction of depression severity. They argue the BDI-II form [38] queries personal circumstances from subjects which cannot be determined from only watching a video with the subject talking. We agree with this observation of Kachele et al. For example, the BDI-II form asks subjects whether they consider themselves as a failure and whether they feel disgusted with themselves. We believe that unless participants explicitly answer these questions, it is a difficult task to predict their response of such questions. For the AVEC 2014 DSC, Kachele et al. used meta-knowledge such as duration of video, movement of subjects as pixel-wise difference of frames, estimated age, gender etc. to achieve an RMSE = 9.58 on the test partition.

Perez et al. [295] investigated the correlation between affective dimensions of arousal, valence, and dominance with depression severity. They experimentally verified that leveraging affective dimensions can be fruitful for automated depression screening, achieving Pearson correlation values of $r = 0.46$ and $r = 0.52$ for Freeform and Northwind tasks, respectively. The best results on the test partition i.e. RMSE = 10.82 are, however, from a brute-force approach (not a model based on affect). This involves combination of audio-visual modalities with feature selection based on relief algorithm [107].

Senoussaoui et al. [154] demonstrated the effectiveness of an i-vector based representation [257] for audio-visual LLDs as part of their solution for the AVEC 2014 DSC. For the audio modality, they used MFCCs along with its velocity and acceleration contours. For the video modality, they used LBP-TOP features [81] which were provided as baseline features. Also, rather than using a single regression model to predict BDI-II scores, they proposed a two-stage model. In the first stage, they classify whether or not a particular subject is depression, then in the second stage, they use a regression model to predict BDI-II scores. The caveat to this approach is that if the classification model fails to predict accurately, the subsequent regression model may also fail catastrophically. The best result they achieved on the test partition was an RMSE = 10.43 which was slightly better than the baseline RMSE = 10.83. It must be mentioned here that they experimented with SVR, generalised linear models (GLM) [296], and relevance vector machine (RVM) [121] for building regression models. They also experimented with a number of combinations of feature level fusion followed by dimensionality reduction using PCA. Thus, it is difficult to distil the effectiveness their methodology beyond RMSE scores.

Mitra et al. [164] explored the usefulness of a diverse set of audio features for predicting BDI-II scores. These features, 18 in total, focus on quantifying various aspects of speech such as: articulatory trajectories, acoustic characteristics, acoustic-phonetic characteristics, and prosodic features. We refer the reader to their paper for details of these features. They evaluated the performance of these features using LOOCV on the development partition, and report that the RMSE of individual features varied between 9.18 and 11.87. On the test partition, their best system achieved RMSE = 11.10. It is reminded that the baseline for audio modality is RMSE = 12.57, and for the video modality, the baseline is RMSE = 10.86. Therefore, using speech alone, they beat the baseline.

Jain et al. [297] perform experiments with Fisher Vector encoding of audio-visual features for the task of depression severity prediction. For the visual modality, they used LBP-TOP [81] and dense trajectories [298]. For audio modality, they used baseline audio features provided as part of the AVEC 2014 DSC dataset. Descriptors from audio and visual modalities are encoded using Fisher Vector encoding [98]. Their motivation for using Fisher vector encoding method was based on state-of-art performance achieved on image classification tasks using this approach. They compare the RMSE of LBP-TOP, dense trajectories, and audio features explicitly and then as feature-level fusion. Amongst these LBP-TOP features encoded as Fisher Vectors achieve the best result on the development partition i.e. RMSE = 8.17, which is better than the baseline RMSE = 9.26 for the development partition. Since there was no improvement in the RMSE due to feature level fusion of additional features, on the test partition they only used LBP-TOP features and achieved an RMSE = 10.25, which is better than baseline RMSE = 10.86.

Jan et al. [299] surmised that depression causes subtle changes in facial and

vocal expression. In order to capture these dynamics, they used motion history histograms (MHH) [300] to quantify local dynamics from audio and visual modalities. For the visual modality, they computed histogram of oriented gradients (HOG) [78], local phase quantization (LPQ) [301], and local binary pattern (LBP) [81] features for each frame of the video. This was followed by using 2D MHH for capturing intra-frame variations in these features.

For audio modality, they used baseline features provided as part of the dataset. Since there were 2,268 features, they tested the performance of each individual feature on the development partition and selected top eight features. Jan et al. do not provide details of the method they used to test the performance of audio features. Next, they experimented with 256 (2^8) combinations of the top eight features. The combination of features which achieved best results include: spectral flatness, Band1000, psychoacoustic sharpness, probability of voicing, shimmer, and zero crossing rate. It is interesting to note that the final selection of features did not feature MFCC features, even though they have previously been shown to be useful for the task of depression screening [162]. Finally, they use one-dimensional MHH to summarise dynamics of the select audio features.

They used PLS regression [103] for predicting BDI-II from audio and visual features, and follow this up with linear regression for fusing predictions from all features. On the test partition, they achieve best results of RMSE = 10.26 by combining audio-visual modalities, which is better than the baseline results on the test partition i.e. RMSE = 10.86.

The winners of the AVEC 2014 DSC, Williamson et al. [128] base their work on the theory that neuro-physiological changes occurring due to depression alter motor control, thereby affecting the mechanisms controlling speech production and facial expression production. They use correlation structure features for quantifying these deficits. When used with formants and the velocity of MFCC contours, the authors find that these coupling features are particularly useful for detecting depression. Williamson et al. call these features vocal tract coordination (VTC) features.

We believe that the key to this method actually lies in measuring the coupling strength of time series because this approach has been successfully used for a variety of tasks such as prediction of seizures from electroencephalogram signals [302], cognitive impairment prediction in the elderly [185, 262], and estimating load carriage from sensors placed on the body [303].

The second vital component of their model is the use of a GMM-based multivariate regression scheme which they call Gaussian staircase regression (GSR). Williamson et al. had proposed GSR previously in [210]. In this approach, training data is split in terms of ordinal ranges of the target variable for regression thereby forming a 'staircase', and an ensemble of Gaussian classifiers is trained to classify over the staircase. For the case of AVEC 2014 DSC, the regression variable is the BDI-II score. We recommend the reader to experimental work of Cummins et al. [246] for further details on

GSR. Williamson et al. computed correlation structure features for formant–CPP, CPP–HNR, and delta MFCC for the Northwind task (which is scripted speech) from audio modality, and Facial Action Units (FAUs) from the visual modality. FAUs were computed using the Computer Expression Recognition Toolbox (CERT) [84]. One must note that while formants and MFCCs are multi-dimensional features, correlation structure features are computed for scalar features only. The authors do not provide details about exactly which features were used to compute correlation structure features. Nevertheless, they achieve an RMSE = 8.12 on the test partition, which beats the baseline score RMSE = 10.26.

4.6.2 AVEC 2016 DCC

We note that none of papers submitted to AVEC 2016 DCC beat the challenge baseline significantly. As pointed out by Williamson et al. [130], the challenge baseline was set with a different ratio of depressed/non-depressed classes as compared to what was provided to participants of the challenge. Therefore, a direct comparison with the baseline may not be fair but one has no choice but to go ahead with this comparison since dataset partitioning is beyond our control.

The only paper to beat the test partition baseline for the AVEC 2016 DCC is the work of Yang et.al. [149]. Their method achieves a mean F1 score of 0.72 compared to 0.70 of the challenge baseline. From visual modality, they used geometric features computed from facial landmarks and HOG features which were provided along with the dataset by challenge organisers. Geometric features are computed in the form of Euclidean distance between pairs of 51 stable 2D facial landmarks, and HOG features are used after dimensionality reduction using PCA. Multiple SVRs with RBF kernels were trained on the training and development partitions and the one which provided the best performance was used for the test partition. From audio modality they also used COVAREP and formant features after reducing dimensionality using PCA. The best result they achieved is using the text modality. Here, they manually selected information from interview transcripts which was related to the PHQ form and used decision trees to classify between depressed and non-depressed classes.

Nasir et al. [152] achieved a mean F1 score of 0.55 on the test partition, which is below the challenge baseline of 0.70. They conducted a large number of experiments for audio/visual features as well as feature selection and classification methods. We believe that the most important contribution from their work is the proposal to use multi-resolution windows to combine LLD from interview sessions. They argue that depression cannot be recognised at temporal resolution of LLD features i.e. 10 ms for audio LLDs and 33 ms for visual LLDs [12].

Williamson et al. [130], the winners of the AVEC 2014 DSC, used audio, visual, and textual modalities as part of their solution for the AVEC 2016 DCC. For audio/visual modalities, their work is similar to their approach for AVEC 2014 DSC. They used textual analysis to mine for answers to questions relevant to PHQ questionnaire, similar to the approach of Yang et al. [149] and also find those features to be most useful. Williamson et al. determined via experimentation on the training and development partition that text modality is much better than any other modality for the task at hand i.e. text modality achieved a maximum mean F1 score of 0.81, whereas audio and visual modalities achieved a maximum mean F1 scores of 0.76 and 0.53, respectively. On the test partition, they achieved score of 0.70, which exactly matches the challenge baseline. Interestingly, the highest accuracy they achieved was from textual analysis of Ellie's script, rather than textual analysis of subject's script.

The submission of Pampouchidou et al. [150] did not beat the baseline i.e. they achieved a score of 0.66 compared to 0.70 of the challenge baseline. Similar to [304], they used the concept of motion history images (MHI) [208]. Since facial images were not provided as part of the dataset due to ethical restrictions, they computed MHI on facial landmarks. Prior to computing MHI, facial landmarks were registered with respect to facial landmarks on the temples, chin, and inner and outer corners of the eyes. They computed LBP and HOG features for each MHI image and used these features for classification.

Huang et al. [155] investigated the effect of subject's gender as a confounding factor for depression recognition for the AVEC 2016 DCC and report that gender did not influence their results. They used relevance vector machines (RVM) [121] which employ a Bayesian framework for regression and classification tasks. Huang et al. also attempted the use of correlation features of time series as well as staircase regression approach, previously used by Williamson et al. [130]. However, they failed to replicate good performance. In fact, the best results they achieve is a score of 0.55 compared to the baseline of 0.70.

Ma et al. [305] provide the only entry in the challenge to use aspects of deep learning. They call their model *DepAudioNet* for audio based depression recognition. Ma et al. propose a combination of convolutional neural networks (CNN) and long short term memory (LSTM) for audio feature representation. While their approach beats the baseline for audio modality i.e. 0.61 versus 0.50 on the development partition, the authors do not provide results for the test partition.

4.6.3 AVEC 2017 DSC

Yang et al. [92,306] submitted two papers for AVEC 2017 DCC Workshop, both were essentially similar in that they estimated depression severity from audio, video, and text modalities using a combination of CNN and fully connected deep neural network (DNN).

For audio features, they computed exhaustive audio descriptors from openSmile toolkit [30] i.e. 238 LLDs, then used 29 functionals to summarise these features, which results in 6,902 features. These features were fed to a CNN to learn deep learnt features.

For text modality, they analysed subjects' answers to questions related to psychoanalytic aspects associated to depression symptoms. These included topics related to: (1) Prior depression diagnoses, (2) Prior post-traumatic stress disorder (PTSD) diagnosis, (3) Sleep disorder, (4) Feeling, and (5) Personality. Next, they used Paragraph Vector (PV) descriptors, introduced by Quoc et al. in [307] to represent sentences for interview transcript.

In one of their papers i.e. [306], Yang et al. proposed Histogram of Displacement Range (HDR) to summarise the movement of facial landmarks. HDR is essentially a histogram of changes in pixel coordinates of facial landmarks. Meanwhile, in their second paper [92], Yang et al. used Motion History Histogram (MHH) [300] to summarise information from AU features.

To deal with dataset imbalance, they divided each recording to multiple segments and consider those segments as new samples. They used a 3 layer CNN each for feature extraction from audio, visual, and text modality. Features learnt from CNNs for the three modalities are fused using a 4 layer DNN. For their paper [92], they report an RMSE = 5.79 on the test partition, whereas for their paper [306], they achieved an RMSE = 5.40. Both of these are better than the baseline RMSE = 6.97 on the test partition.

Dang et al. [158] investigated the effectiveness of word affect for predicting depression severity. They base their work on prior research which has shown that depressed individuals exhibit more negative sentiments compared to non-depressed individuals [308]. They used the Suite of Automatic Linguistic Analysis Tools (SALAT) [309] for extracting a range of text-based features from transcripts of subjects' interviews with the virtual agent.

Apart from the text modality, they also experimented with audio and visual modalities. For the audio modality, they used MFCC features, whereas for the video modality they used FAUs. For regression, they used Gaussian staircase regression [210].

While they experimented on audio, visual, and text modalities for predicting depression severity, their best performance of RMSE = 6.02 on the test partition was achieved whilst using Affective Norms for English Words (ANEW) features [310]. ANEW features provide affective normal for English words in terms of arousal, valence, dominance, and pleasure. It is important to mention here that RMSE on the test partition decreased when they appended other text or audio/visual features with ANEW features.

Sun et al. [153] report on experiments with audio/visual modalities for predicting depression severity. Their first approach is essentially brute-force where they used ensembles of cascaded random forest regressors, with a random forest for each baseline paper (see Section 4.5.2). In their second method, which uses the text modality, Sun et al. manually selected features from interview

transcripts. They specifically focussed on subject's answers to questions related to *previous PTSD/depression diagnoses*, *treatment of mental illness*, *personal preference and feelings*, and *sleep quality*. We note that these questions are similar to those asked on the PHQ-8 [21] questionnaire, which is exactly what makes up depression severity scores.

On the test partition, their brute-force approach with cascades of random forests achieved RMSE = 6.22, which is better than the baseline for video modality i.e. RMSE = 6.97. However, their best result on the test partition is achieved with manually selected text features, with which they achieve an RMSE = 4.98.

Gong et al. [131] were the winners of the AVEC 2017 DCC via their topic modelling based approach to perform context-aware analysis of the audio-visual recordings. They surmised that since Ellie is a virtual agent, therefore, it will have a limited number of topics which it can use as part of the interview process. Furthermore, the number of topics possible in clinical interviews are also limited (for example, the topics likely focus on PHQ-8 questionnaire for the AVEC 2017 DCC dataset). They start by building a sentence dictionary based on the questions asked by Ellie, and manually clean the sentence dictionary of inconsistencies. This process yields 83 topics (see Table 1 of [131] for the list of topics).

Next, for topic they each create a feature vector which contains subject's gender, FAUs (mean, max, and min functionals), COVAREP features (mean, max, and min functionals), and semantic features based on linguistic inquiry and word count (LIWC) [311], that is 390 features for each topic. In total, for 83 topics, the length of feature vector reaches a length of 32,370, but this feature vector contains audio, video, and text information of each topic. In order reduce the dimensionality of the feature vector, they use correlation-based feature subset selection (CFS) [312].

Finally for predicting depression severity, they experimented with decision trees, random forests, SVR, and Stochastic Gradient Descent (SGD) regression using the development partition. Amongst these regression methods, they found SGD regression to achieve best results. On the test partition, the achieve RMSE = 4.99, which is considerably better than the best baseline for AVEC 2016 DCC i.e. RMSE = 6.97.

The reader may note that Sun et al. [153] achieved an RMSE = 4.98 on the test partition, whereas Gong et al. [131] achieved an RMSE = 4.99 — which were deemed within the margin of error by the organisers of the event. However, since Sun et al. achieved their result using manually selected text features, Gong et al. were declared winners of the AVEC 2016 DCC .

4.7 Methodology

In this section, we discuss the methodology of our multi-pronged approach for the development of automated screening of depression.

4.7.1 Turbulence in Feature Contours

We discussed in Section 3.4 that psychomotor symptoms lead to uniqueness in an individual's speech pattern, here we posit that it must manifest itself as turbulence or lack thereof in speech feature trajectories. Fundamentally, we hypothesise the LLDs of speech of individuals with and without depression are different, and if quantified may provide an insight into depression severity. We call these features as *turbulence features*.

However, given the complex nature of depression and how it affects the speech production system, the task of recognising turbulent patterns in speech is complicated. Inspired by the method of formulation of the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) [58], we devise the following methodology to capture the hypothesised turbulence, and later demonstrate its effectiveness for the task of depression screening.

Consider as an example, the pitch contour of an individual's speech (F0 feature). It has been computed at a frequency of 100 Hz using the COVAREP toolbox [62]. Due to the free speech nature of the recordings in the dataset, there exists no prior knowledge where this turbulence may manifest. We therefore consider a multi-scale approach, by using a set of temporal windows of lengths $\{0.5, 2, 5, 10, 15\}$ seconds, with an overlap of $\{0.2, 1, 3, 5, 7\}$ seconds, respectively. Within each window, we compute the crest factor as the measure of turbulence [313]. The crest factor measures the ratio between the absolute max value of the signal and its root mean square (RMS) value. Therefore, if there indeed exist irregularities in the pitch within any window, we are likely to capture them. Finally, the crest factor values from multiple windows at each scale are pooled using the following descriptive statistics: the 10^{th}, 25^{th}, 50^{th}, and 75^{th} percentile, and the mean with 5% trimming.

In addition to the pitch, we apply the above multi-scale procedure to a number of voice quality features. Our motivation is to capture changes in voice quality due to depression. We compute voice quality features using COVAREP toolbox [62]. These features include: normalised amplitude quotient (NAQ), quasi open quotient (QOQ), the difference in amplitude of the first two harmonics of the differentiated glottal source spectrum (H1H2), parabolic spectral parameter (PSP), maxima dispersion quotient (MDQ), spectral tilt/slope of wavelet responses (PS), and shape parameter of the Liljencrants-Fantmodel of the glottal pulse dynamics (Rd). These features were computed using the COVAREP toolkit [31].

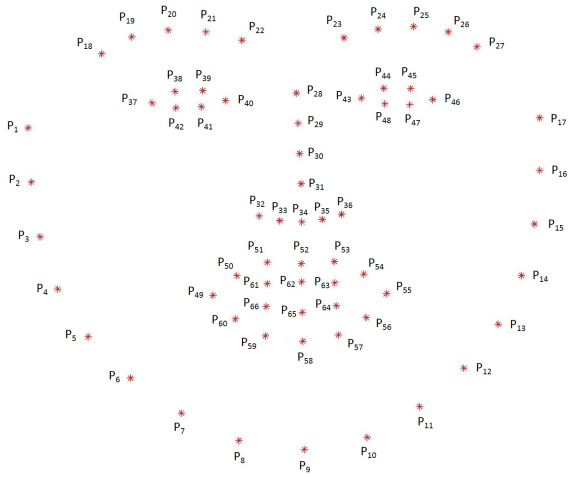


Figure 4.1: Reference for numbering of 66 point facial landmarks

Nasir et al. [152] have argued that depression is a long term effect and may not be evident at fine temporal resolutions of LLDs i.e. 10 ms for speech and 33 ms for visual features in our case. In their work, they use a temporal window of 10 s in duration with a shift of 2 s. We concur and use four temporal resolutions i.e. Res 1: 2 s with 1 s shift, Res 2: 5 s with 2 s shift, Res 3: 10 s with 5 s shift and Res 4: 15 s with 7 s shift.

However, in contrast to [152], where features are combined within a particular temporal window using arithmetic mean, we use the following functionals of descriptive statistics: arithmetic mean, 10% trimmed arithmetic mean, standard deviation, median and range (as the difference between the 99^{th} and 1^{st} percentile). We also compute the crest factor, which measures the peak value of the feature in a particular window with respect to the RMS value of that feature in that window. Therefore, it inherently measures rapid changes in a feature value, which we hypothesised will prove to be useful as a descriptor for psychomotor symptoms. We discuss the effectiveness of this approach for the task of automated screening of depression severity in Section 4.8.2.

4.7.2 Modelling Craniofacial Movement

In this section, we discuss our method to model craniofacial muscle movement and later demonstrate the effectiveness of this method on the AVEC 2014 DSC and AVEC 2016 DSC datasets.

Our objective is to craft visual features that are capable of representing muscular tightening, a trait of psychomotor retardation. Thus, we hypothesise that if an individual has depression then their head movement as well as facial muscle movement will be impaired compared to those who do not have depression. In order to track these movements, we use 66-point 3D facial

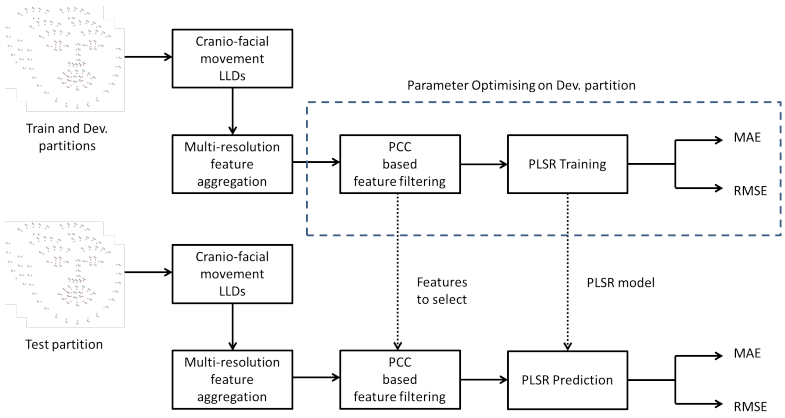


Figure 4.2: Framework for our proposed craniofacial movement features for predicting depression severity score

landmarks, as illustrated in Figure 4.1, of subjects in the two datasets. These facial landmarks are already provided as part of the AVEC 2017 DSC by the organisers of the challenge (ethics restrictions meant that organisers could not provide video recordings [13]). We computed facial landmarks for video recordings provided as part of the AVEC 2014 DSC dataset using the OpenFace toolkit [32] with its default settings i.e. the same software toolkit used to compute facial landmarks for the AVEC 2017 DSC dataset.

Similar to previous work on this subject [175, 207], we compute velocity and acceleration contours from facial landmarks. However, unlike [175, 207] where contours are computed for many combinations of facial landmarks, we specifically target four types of movements: (1) head movement, (2) mouth movement (both horizontal and vertical), (3) eyelid movement, and (4) Eyebrow movement. Details of craniofacial movement features are described in subsequent sections.

In order to demonstrate the efficacy of our proposed approach for computing craniofacial movement, we propose the framework illustrated in Figure 4.2. The first part of the framework focuses on feature generation. Here, we start by computing velocity and acceleration contours of facial landmarks, using these contours to represent craniofacial movement (as detailed in subsequent sections). Next, we use the concept of multi-resolution feature aggregation, as used for turbulence features in Section 4.7.1, to summarise information from velocity and acceleration contours into a fixed length feature vector. We use a set of temporal windows of lengths $\{2, 5, 10, 15\}$ seconds, with an overlap of $\{1, 2, 5, 7\}$ seconds, respectively. Within each window, we compute the median, range, and crest factor to summarise velocity and acceleration contours within each window. Finally, we use min, max, median, range functionals to create a global representation for both velocity and acceleration contours for the entire recording. This completes the stage of multi-resolution feature aggregation in

our framework.

The next stage of the framework implements regression. Here, we first use Pearson's correlation coefficient based filtering to arrange features in descending order with respect to their correlation to the depression severity score. These sorted features are provided as input to PLSR regressor [172]. The PLSR regressor has two parameters which are optimised using the development partition. These include (a) the number of features which are provided as input to the regressor, and (b) the number of components for PLSR. The parameters tuned for the development partition are then passed down to be used for the test partition.

We follow the rules for the AVEC 2014 DSC and AVEC 2017 DSC challenges, where participants were instructed to develop their model using the training and development partitions, and were only allowed to use the test partition for measuring the efficacy of their proposed methods i.e. participants were not permitted to use the test partition for training their models. Given that the objective of both AVEC challenges was to predict scores of depression measurement instruments, the accuracy was measured in terms of RMSE.

In subsequent paragraphs, we shall detail our methodology for crafting craniofacial movement features.

Head movement

For quantifying head movement, we aim to select facial landmarks which are representative of rigid movement of face. We surmise that rigid movement of the face is a correlate for head movement. Ideally, one would compute headpose for this purpose, however, that requires information about parameters [104, 175, 207]. Such information is not available for either of the two datasets.

We investigate the use of two sets of facial landmarks to quantify head movement: the first contains landmarks from the nose region and includes landmarks $\{P_{28}, P_{29}, \ldots, P_{36}\}$. The second set of landmarks form the face contour and includes landmarks $\{P_1, P_2, \ldots, P_{17}\}$. Both of these are considered *stable* since they are not effected by non-rigid facial movement [12, 150]. In addition to these two sets of landmarks, we shall also investigate the efficacy of using both of these landmarks together. A summary of these facial landmarks is summarised in Table 4.2.

The procedure for computing head movement features is as follows: for each set of facial landmarks, we first compute the 3D Euclidean distance between contiguous frames, which encodes the change in (x,y,z) coordinates. Following this procedure, for all frames, generates a vector representing velocity of head movement for the individual. Similarly, applying the second order difference operation to the velocity contour provides the acceleration contour. It is important to mention here that we only use landmarks for which the tracker outputs a success in order to avoid using landmarks for which the tracker has low confidence. The velocity and acceleration contours are then passed to the

Table 4.2: Summary of facial landmarks used for the proposed cranio-facial movement features

Features	**Facial Landmarks**
	Head Movement
Face Contour	$\{P_1, P_2, \ldots, P_{17}\}$
Nose	$\{P_{28}, P_{29}, \ldots, P_{36}\}$
Nose + Face Contour	$\{P_1, P_2, \ldots, P_{17}, P_{28}, P_{29}, \ldots, P_{36}\}$
	Mouth Movement
HorizCentral	$\{\|P_{51}P_{53}\|, \|P_{57}P_{59}\|, \|P_{61}P_{63}\|, \|P_{64}P_{66}\|\}$
HorizNonCentral	$\{\|P_{50}P_{54}\|, \|P_{49}P_{55}\|, \|P_{56}P_{60}\|\}$
VertCentral	$\{\|P_{52}P_{58}\|$
VertNonCentral	$\{\|P_{50}P_{60}\|, \|P_{51}P_{59}\|, \|P_{53}P_{57}\|\}, \|P_{54}P_{56}\|\}$
DiagCentral	$\{\|P_{51}P_{57}\|, \|P_{53}P_{59}\|\}$
DiagNonCentral	$\{\|P_{50}P_{56}\|, \|P_{54}P_{60}\|\}$
	Eyelid Movement
Vert	Left: $\{\|P_{44}P_{48}\|, \|P_{45}P_{47}\|\}$ Right: $\{\|P_{38}P_{42}\|, \|P_{39}P_{41}\|\}$
Diag	Left: $\{\|P_{44}P_{47}\|, \|P_{45}P_{48}\|\}$ Right: $\{\|P_{38}P_{41}\|, \|P_{39}P_{42}\|\}$
	Eyebrow Movement
θ_1	Left: $\triangle P_{23}P_{24}P_{43}$ Right: $\triangle P_{21}P_{22}P_{40}$
θ_2	Left: $\triangle P_{23}P_{25}P_{43}$ Right: $\triangle P_{20}P_{22}P_{40}$
θ_3	Left: $\triangle P_{23}P_{26}P_{43}$ Right: $\triangle P_{19}P_{22}P_{40}$
θ_4	Left: $\triangle P_{23}P_{27}P_{43}$ Right: $\triangle P_{18}P_{22}P_{40}$

multi-resolution feature aggregation and regression framework as illustrated in Figure 4.2.

Mouth Movement

These features quantify deformations of the mouth region as the subject speaks to the camera. It is surmised that by modelling mouth movements, one can jointly represent the amount of speech by the subject as well as the amount of effort the subject puts in order to produce it. For the sake of completeness, we first quantify horizontal, vertical, and diagonal deformations of the mouth

separately and later investigate the modelling of all deformations together.

For quantifying horizontal deformations of the mouth, the first step is to compute, for every frame, the pairwise Euclidean distance between the landmarks P_{49} and P_{55}, P_{50} and P_{54}, P_{60} and P_{56}, P_{61} and P_{63}, P_{66} and P_{64}, P_{51} and P_{53}, and P_{59} and P_{57}, representing them as $||P_{49}P_{55}||$, $||P_{50}P_{54}||$, $||P_{60}P_{56}||$, $\|P_{61}P_{63}\|$, $\|P_{66}P_{64}\|$, $\|P_{51}P_{53}\|$, and $\|P_{59}P_{57}\|$, respectively, as illustrated in Figure 4.3.

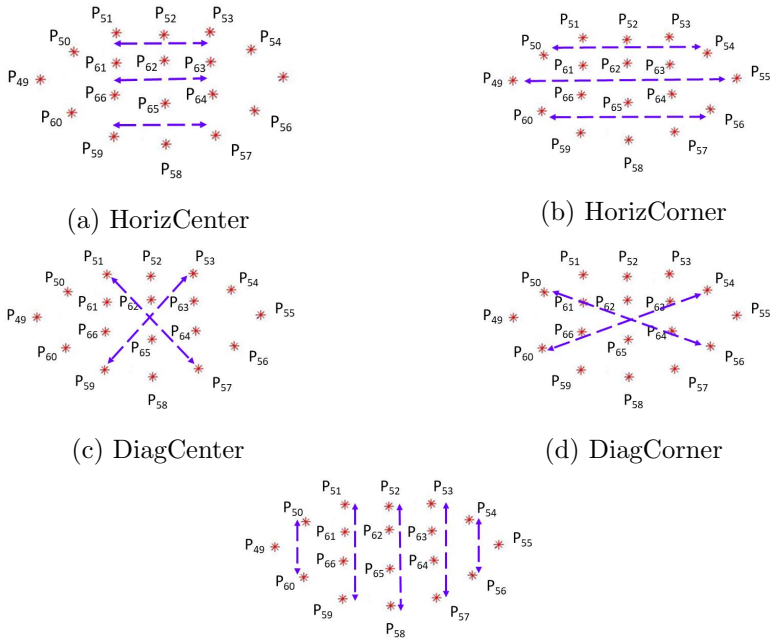
While it is possible to use all the mentioned pairs of facial landmarks for quantifying horizontal deformations, our aim is to determine the optimal landmarks for the task at hand. We do not want to use more landmarks than what are necessary, but also do not want to use a brute force approach such as that based on feature selection [100]. Hence, we organise these distance measures into two sets as summarised in Table 4.2. The *HorizCentral* set consists of landmarks closer to centre of the mouth region, whereas the *HorizNonCentral* set consists of landmarks closer to lip corners. We surmise that landmarks closer to the lip corners may be more relevant since they experience greater change during smiling compared to landmarks closer to the center of the mouth region. The reader is reminded that individuals with depression tend to smile less or have unnatural smiles [180, 192]. Next, for each of the two sets of landmarks, velocity and acceleration contours are computed by applying 1st and 2nd order difference operators. Finally, the average value of each contour is taken for landmarks in *HorizCentral* and *HorizNonCentral* contours to yield horizontal mouth movement features for the subject. These features are then passed down to the multi-scale feature aggregation stage of our framework.

For vertical movement, we follow a similar procedure to that used for quantification of horizontal movement. For every frame, we compute pairwise distances between facial landmarks i.e. $\|P_{51}P_{59}\|$, $\|P_{52}P_{58}\|$, $\|P_{53}P_{57}\|$, $\|P_{61}P_{66}\|$, $\|P_{62}P_{65}\|$, and $\|P_{63}P_{64}\|$, as illustrated in Figure 4.3. Next, these pairwise distance measures are organised into two sets i.e. *VertCentral* and *VertNonCentral* based on whether the landmarks are closer to the centre of the mouth region or the lip corners. The rest of the process is same as followed for horizontal movement features.

Finally, we compute craniofacial movement features which represent diagonal deformations of the mouth. As summarised in Table 4.2, we compute *DiagCentral* and *DiagNonCentral* sets of features to quantify diagonal mouth deformations. The rest of the process for diagonal mouth movement is same as that used for horizontal and vertical features.

Eyelid movements

We measure eyelid movement as a correlate of blinking rate, which according to [174, 314, 315] can be used to identify individuals who have depression. We investigate two types of features for quantifying eyelid movements, as summarised in Table 4.2. The first set of eyelid movement features is called



(a) HorizCenter

(b) HorizCorner

(c) DiagCenter

(d) DiagCorner

(e) Vert

Figure 4.3: Reference for facial landmarks from the mouth region used for quantifying mouth movement

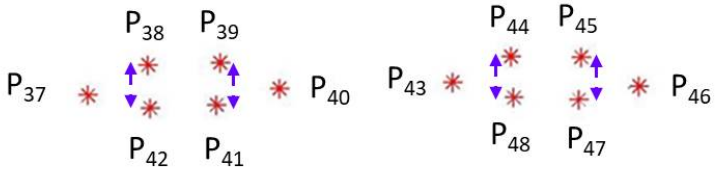
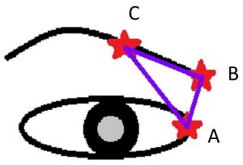
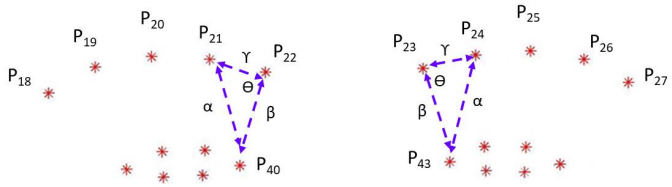


Figure 4.4: Reference for facial landmarks from the eye region used for quantifying Eyelid movement, right eye (left) and left eye (right)

Vert and contains velocity and acceleration contours based on pairwise distances $||P_{38}P_{42}||$ and $||P_{39}P_{41}||$ for the right eye and $||P_{44}P_{48}||$ and $||P_{45}P_{47}||$ for the left eye, as illustrated in Figure 4.4. Meanwhile, the second set of eyelid movement features is called *Diag*, and as the name suggests it contains velocity and acceleration contours for diagonal distance to characterise eye opening. While we compute the *Diag* features for the sake of thorough investigation, one can argue that *Vert* features provide a more direct method to quantify eye opening.



(a) Illustration of method to compute eyebrow movement



(b) Reference for facial landmarks from the eye and eyebrow region used for quantifying eyebrow movement

Figure 4.5: Quantifying Eyebrow movement

Eyebrow movements

Sobin et al. [202] reported that individuals with depression have greater eyebrow movement. However, their work was based on manual inspection of eyebrow movement in video recordings. Our aim here is to automate this process by crafting features to quantify eyebrow movement using facial landmarks.

In order to quantify eyebrow movement, we propose to measure the angle formed between the eye corner landmark and various landmarks on the eyebrow, as illustrated in Figure 4.5a. Here, landmarks at positions A and B are used as anchor points, with A being the landmark for the eye-corner and B, the landmark for the eyebrow corner. The angle θ formed for the triangle $\triangle ABC$ then provides a measure of the changes due to eyebrow movement. This angle is computed using the Law of Cosines as given in equation 4.1, where α, β and γ are the Euclidean distances between landmarks A and C, B and C, and A and B, respectively.

$$\theta = \arccos\left(\frac{\beta^2 + \gamma^2 - \alpha^2}{2\beta\gamma}\right) \tag{4.1}$$

Given that there are four more eyebrow landmarks (other than the anchor landmark; refer Figure 4.5b, we iteratively select each of the four available options in order to compute angles $\{\theta_1, \theta_2, \theta_3, \theta_4\}$ and later conduct an inves-

tigation in order to determine the best configuration for eyebrow movement features.

As an example, consider Figure 4.5b where the angle $\angle P_{21}P_{22}P_{40}$ needs to be computed to quantify the degree of eyebrow movement for the right eye. The angle $\angle P_{21}P_{22}P_{40}$ is computed using the Law of Cosines, as given in equation 4.1, where α, β, and γ are the Euclidean distances between landmarks P_{21} and P_{40}, P_{22} and P_{40}, and P_{21} and P_{22}, respectively. Here, landmarks P_{22} and P_{40} are used as anchor points, whereas the landmark P_{21} is one of the four landmarks over which we compute the angle θ, as summarised in Table 4.2. A similar process is followed to compute the four angles for left eyebrow as well. Finally, velocity and acceleration contours are computed and the average of velocity and acceleration contours for each angle for both eyes is taken as the feature descriptor for eyebrow movement. These features are subsequently passed down to the multi-resolution feature aggregation stage of our framework.

4.7.3 Fisher Vector encoding of Speech Spectra

We hypothesise that speech spectra of individuals with depression, represented as spectral LLDs, is characteristically different from those who do not suffer from depression, and can be quantified. We propose a framework based on Fisher Vector encoding to test this hypothesis.

In the following sub-sections, we first provide an introduction to Fisher Vector encoding before proceeding to a discussion on our proposed framework for modelling speech spectra using Fisher Vectors.

Fisher Vector Encoding

Fisher Vector encoding is a feature aggregation method which was proposed by Perronnin et al. [36, 99] for use in object recognition tasks, where it provided state-of-the-art results [98] until the advent of the age of deep learning [87].

We note that Fisher Vector encoding also has gained recognition amongst researchers from SSP/AC community. Simonyan et al. [316] achieved state-of-the-art results for the task of face recognition using this approach. Jain et al. [297] used FV encoding of audio baseline features (functionals) for the AVEC 2014 DSC, whereas Dhall et al. [110] used FV encoding of visual features (LBP-TOP) for the task of depression recognition. Meanwhile, Kaya et al. [317] used FV for emotion recognition, and later used it for Computational Paralinguistics [112, 318]. The results reported by these researchers motivated us to investigate the efficacy of Fisher Vector based aggregation of spectral LLDs for modelling speech of individuals with depression.

Fisher Vector encoding owes its success to prior work by Jaakkola and Haussler [319] who proposed the *Fisher Kernel* framework. Their aim was to develop a framework which combined the advantages of generative models (i.e.

ability to work with variable length data items) and discriminative models (i.e. ability to learn class specific boundaries). Jaakkola and Haussler achieved this through kernel functions, computed from generative probabilistic models, which provide discriminative features. The underlying idea, based on Fisher scores, is that within a generative model built using various data items, similarly structured data items will induce similar gradients in the parameters of the generative model. These gradients can subsequently be used as feature vectors for classification tasks. Therefore, it follows that the Fisher Kernel framework provides features based on the underlying probability distribution that are *naturally* discriminative.

Perronin et al. [36, 99] proposed to follow the Fisher Kernel framework for the task of object recognition. The input to the framework are visual descriptors of objects and these represent data items in the Fisher Kernel framework. The generative model is a Gaussian Mixture Model (GMM) [95] which is built using visual descriptors to approximate the distribution of visual descriptors in the entire dataset (thus, acting as the visual vocabulary). The output of their framework are *Fisher Vectors*, which represent the first and second order statistics for gradients between the visual descriptors for each object and the visual vocabulary.

In mathematical form [36, 99], the process can be summarised as follows: Let $I = x_1, x_2, ..., x_N$ be a set of D dimensional feature vectors extracted from an image which represents an object. Let $\theta = (\mu_k, \Sigma_k, \pi_k)\ \forall\ k = 1, 2, ..., K$ be the sufficient parameters of the generative model created using GMM with K Gaussians. The GMM associates each vector x_i to the Gaussian k within the GMM with a strength given by the posterior probability as:

$$q_{i,k} = \frac{exp\left[-\frac{1}{2}\left(x_i - \mu_j\right)^T \Sigma_k^{-1}\left(x_i - \mu_j\right)\right]}{\sum_{t=1}^{K} exp\left[-\frac{1}{2}\left(x_i - \mu_t\right)^T \Sigma_k^{-1}\left(x_i - \mu_t\right)\right]} \tag{4.2}$$

For each Gaussian within the GMM, the mean gradient vector u_k and the covariance gradient vector v_k can be computed as follows:

$$u_{j,k} = \frac{1}{N\sqrt{\pi_k}} \sum_{i=1}^{N} q_{i,k} \frac{x_{j,i} - \mu_{j,k}}{\sigma_{j,k}} \tag{4.3}$$

$$v_{j,k} = \frac{1}{N\sqrt{2\pi_k}} \sum_{i=1}^{N} q_{i,k} \left[\left(\frac{x_{j,i} - \mu_{j,k}}{\sigma_{j,k}}\right)^2 - 1\right] \tag{4.4}$$

Note that the mean and covariance gradient vectors span the $j = 1, 2, ..., D$ vector dimensions of the input feature descriptor. Finally, the Fisher Vector representing the image is achieved by concatenating the two gradient vectors, i.e.

$$\Phi = [u_{1,k}, u_{2,k}, ..., u_{D,k}, v_{1,k}, v_{2,k}, ..., v_{D,k}] \tag{4.5}$$

Perronin et al.'s Fisher Vectors come in two versions. The first version is commonly known as *vanilla* Fisher Vectors and represents the original formulation [36]. The second version is called *improved* Fisher Vectors and was proposed in [99] i.e. three years after their original formulation. In *improved* Fisher Vectors, Perronin et al. proposed methods to normalise Fisher Vectors in a two stage process. They demonstrated that normalisation of Fisher Vectors leads to improvement in classification accuracy for object detection task when compared with results based on *vanilla* Fisher Vectors.

They proposed *power normalisation* to deal with cases where the input feature descriptors are not independent (an assumption in their framework). The efficacy of their proposed normalisation was demonstrated in [98, 99]. Perronin et al. also proposed *L2-normalisation* based on the argument that it is always beneficial for high-dimensional feature vectors when used in combination with linear classifiers. The efficacy of L2-normalisation was demonstrated in [98, 99, 320].

Proposed Framework for Depression Recognition

In order to model speech spectra of individuals with depression using Fisher Vectors, we first compute various spectral LLDs. These include Mel Frequency Cepstral Coefficients (MFCCs) and Perceptual Linear Prediction (PLP) coefficients, which are standard representations for spectra in speech processing. These LLDS, along with their velocity and acceleration contours were computed using the openSmile toolkit version 2.3.0 [30]. Apart from MFCCs and PLP coefficients, we also experiment with FV encoding of formants, Mel-Cepstral compression (MCEP), Harmonic Model plus Phase Distortion Mean (HMDPM) and Harmonic Model plus Phase Distortion Deviation (HMDPD) computed using standard settings of COVAREP toolbox version 1.3.2 [62].

The overall layout of our framework is depicted in Figure 4.6: we start by concatenating spectral LLDs from each speech recording of the training partition into a single matrix and then build a background model for the spectral space using a GMM.

However, in order to train the GMM efficiently, we perform the following pre-processing steps: all feature frames are z-score normalised i.e. made to have zero mean and unit standard deviation. Next, we use principal component analysis (PCA) to decorrelate the feature space. We retain dimensions such that they match the number of clusters of the GMM. In case the feature set has dimensionality smaller than or equal to the number of clusters for GMM, we perform PCA over the dimensions of feature set, i.e. dimensionality reduction does not take place. We apply a second z-score normalisation on the output of PCA before using the resultant features to fit the GMM. Finally, we compute Fisher Vectors in order to describe the deviation of each participant's spectra from the background model.

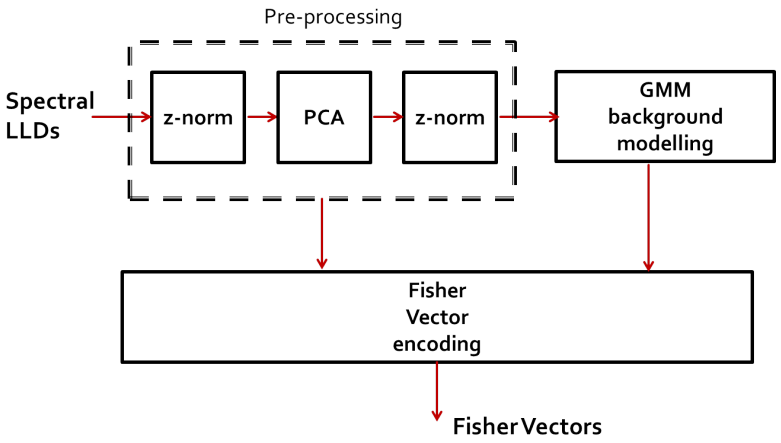


Figure 4.6: Block diagram for FV encoding of spectral LLDs.

It is also important to mention here that since the GMM is built using pre-processed features, the combination of z-norm + PCA + z-norm needs to be applied to any new features which need to be encoded as Fisher Vectors, for example, features of individuals from the development and test partitions.

We used VLFeat library [321] for both, estimating the means, covariance matrix, and priors of the GMM, and implementing FV encoding. For FV encoding we set parameters such that Perronnin's improved FVs [99] are computed.

4.7.4 Weighted Extreme Learning Machines

We use Weighted Extreme Learning Machines (WELMs) [322] for classifying speech of individuals with depression and healthy individuals. Our motivation for using WELMs is based on the work of Williamson et al. [128] who argued that with limited data, least squares based approach would perform better than stochastic gradient descent approach of SVM. Here, we provide a brief introduction to WELMs.

ELMs is essentially a single layer feed-forward neural network where the hidden layer is assigned randomly generated weights and these weights are not updated during the training process. The classifier works by mapping the output from the hidden layer to the training labels using a least squares fit [122, 123]. The idea is that even with random weights, the hidden layer can learn useful representation of input data which can be exploited by designing a suitable output layer. An outstanding advantage of ELMs is their very fast training time, which eases the process of tuning its hyper-parameters and further experimentation.

We note, from the training and the development partitions of the AVEC 2016 DCC, that about 80% of participants are labelled as non-depressed while

the remaining 20% are labelled as depressed. Due to this class imbalance, we opt to use WELMs as proposed in [322] rather than legacy ELM (non-weighted) classifier. WELMs assign weights to each class according to the number of training examples available for that class. The WELM classifier has three hyper-parameters: (1) the number of neurons L in the hidden layer, (2) the regularisation parameter c required for the generalised Moore-Penrose inverse [322], and (3) the weights. While all of these parameters can be tuned using grid search, we determine weights based on the class distribution in the training partition.

The mathematical formulation for WELMs is given as follows [322]: Let H, an $N \times M$ matrix, be the output of the hidden layer (i.e. after multiplication of feature matrix with randomly generated weights of the hidden layer), with N samples and M features, $T \in \{1, 2\}$ be an $N \times 1$ vector which contains the labels for an individual not having depression ($T = 1$) and having depression ($T = 2$), and β be the transformation (aka output weights) which maps H to T, given as $H\beta = T$. Then according to the generalised Moore-Penrose generalised inverse for H, the transformation can be computed either using eq. 4.6 or eq. 4.7, as follows:

when $N \leq L$

$$\beta = H^T \left(\frac{I}{c} + WHH^T \right)^{-1} \tag{4.6}$$

when $N > L$

$$\beta = \left(\frac{I}{c} + H^T W H \right)^{-1} H^T W T \tag{4.7}$$

where W are the user defined weights to take into account class imbalance, I is an identity matrix used to integrate the regularisation parameter c in equations 4.6 and 4.7.

4.8 Experimental Results and Discussion

4.8.1 Craniofacial Movement Features

The results for our proposed craniofacial movement features have been summarised in Table 4.3 through Table 4.6. Here, craniofacial movement features were computed for all subjects in two datasets from AVEC 2017 [13] and AVEC 2014 [11] challenges on depressions severity estimation. The objective of these challenges was to make predictions on the score of the depression severity measurement instrument. For the AVEC 2017, PHQ-8 [21] was used as the depression severity measurement instrument, whereas for AVEC 2014 BDI-II [38] was used. As per the rules of AVEC, performance is measured in terms of the RMSE.

Table 4.3: Performance analysis of Head movement features for AVEC 2017 and AVEC 2014 datasets

Features	**RMSE**		**MAE**	
	Dev	Test	Dev	Test
AVEC 2017				
Face	6.43	6.14	5.27	5.26
Nose	6.05	6.27	4.96	5.32
Face & Nose	6.39	6.71	5.13	5.66
AVEC 2014				
Face	11.06	10.66	9.19	8.64
Nose	10.71	10.96	9.22	9.04
Face & Nose	10.73	10.56	9.14	8.62

Table 4.4: Performance analysis of Mouth movement features for AVEC 2017 and AVEC 2014 datasets

Features	**RMSE**		**MAE**	
	Dev	Test	Dev	Test
AVEC 2017				
HorizCentral	6.54	5.99	5.39	5.03
HorizNonCentral	6.05	6.20	4.98	4.94
VertCentral	6.64	6.34	5.48	5.42
VertNonCentral	6.66	6.44	5.40	5.46
DiagCentral	6.56	6.35	5.42	5.36
DiagNonCentral	6.41	6.06	5.21	5.04
Avg. of All	6.49	6.32	5.41	5.29
AVEC 2014				
HorizCentral	10.86	11.59	9.15	9.49
HorizNonCentral	10.12	11.19	8.50	9.37
VertCentral	11.32	13.44	9.44	10.88
VertNonCentral	11.55	12.59	9.58	10.41
DiagCentral	10.93	13.20	9.28	10.65
DiagNonCentral	11.11	11.41	9.21	9.63
Avg. of All	11.00	13.46	9.35	10.64

The results for head movement features are summarised in Table 4.3. Here we investigated the efficacy of using two sets of facial landmarks which are not affected by non-rigid facial movement (such as facial expressions) for the task of quantifying head movement. For AVEC 2017 dataset, we note that the facial landmarks from the nose region provide the best RMSE on the development partition, whereas, landmarks from the face region provide the best RMSE on the test partition. It is not uncommon to have slightly different performances for the development and test partition. This is due to the fact that the two partitions contain different subjects and while organisers aim to create balanced partitions in terms of label distribution, there are aspects, such as, age, gender, and personality, which influence social signals of subjects (refer Section 3.5.5).

However, as per "the rules of the game" [11], the test partition is only to be used for measuring efficacy of proposed methods, therefore, one must select features based on the results of the development partition only. For the AVEC 2014 dataset, we again note that nose based features provide the best results in terms of RMSE, however, it is closely followed by the a combination of face and nose features. Overall, on the basis of the results for the two datasets, we propose the use of nose based features to quantify head movement.

In Table 4.4, we summarise the results for mouth movement features. Here we investigated the efficacy of features which quantify horizontal, vertical, and diagonal deformations of the mouth region. Furthermore, we also investigated whether it is better to craft features from landmarks which are at a central location in the mouth region compared to those which are relatively further away from the central position. Results indicate that *HorizNonCentral* features are best features for quantifying mouth movement, when the objective is to achieve smallest RMSE for the development partition. Meanwhile, we note that *VertCentral* and *VertNonCentral* features, which quantify mouth openings, achieved the worst RMSE on the development partitions for both datasets.

We find it particularly interesting that horizontal movement features have smaller RMSE than vertical mouth movement features. One possible explanation is that horizontal movement features capture movement of the zygomatic major facial muscle i.e. the lip-corner puller. The lip-corner puller is the dominant action unit for *happiness* emotion according to Paul Ekman's proposal for representation of fundamental emotions from facial muscle movements [7]. Horizontal mouth movement is also responsible for production of certain vowels [323], and as reported by Scherer et al. [324], individuals under distress do indeed have a reduced vowel space. As expected *Diag* mouth movement features provide an intermediate performance between *Horiz* and *Vert* features given that diagonal deformation of mouth is based on simultaneous horizontal and vertical deformations.

In Table 4.5, we summarise the results for eyebrow movement features. Here, our aim was to quantify deformations in the eyebrow caused by movement of muscles in the eyebrow region. While our results do not point out a particular

Table 4.5: Performance analysis of Eyebrow movement features for AVEC 2017 and AVEC 2014 datasets

Features	**RMSE**		**MAE**	
	Dev	Test	Dev	Test
	AVEC 2017			
θ_1	6.48	6.33	5.46	5.29
θ_2	6.34	6.34	5.32	5.18
θ_3	6.36	6.38	5.28	5.32
θ_4	6.28	6.36	5.36	5.35
Avg. of All	6.27	6.29	5.26	5.13
	AVEC 2014			
θ_1	11.28	11.90	9.51	9.90
θ_2	11.09	11.89	9.25	9.80
θ_3	11.15	11.79	9.40	9.80
θ_4	11.24	11.56	9.36	9.72
Avg. of All	11.11	11.80	9.28	9.71

feature which stands out for both datasets, we do note that the feature θ_1 provides worst performance. It is reminded that θ_1 is the angle computed between the anchor landmarks and the eyebrow landmark closest to the center of the face.

Given that there is no conclusive evidence which favours a particular eyebrow movement feature, we propose to use the average value of all these features as the eyebrow movement feature. This proposal is backed by results shown in Table 4.5, the average feature achieves the best performance on the development partition for both datasets, as well as achieving the best performance on the test partition for the AVEC 2017 dataset.

In Table 4.6, we summarise the results for eyelid movement features. Here, our aim was to quantify eye openings using facial landmarks in the eye region and subsequently use it for prediction of depression severity. To this end, we investigated the efficacy of *Vert* and *Diag* eyelid movement features for the tasks. Results indicate the *Vert* features provide better results as compared to *Diag* features, which is expected given that the latter is a direct measure of eye openings.

Overall, we conclude on the basis of our investigation for the proposed craniofacial movement features and associated results that head movement, based on features derived from the nose region, is most useful for predicting the depression severity score. This is followed by mouth movement, in particular those features which quantify horizontal deformations. We also report that eyebrow movement features need to consider changes in eyebrow deformations

Table 4.6: Performance analysis of Eyelid movement features for AVEC 2017 and AVEC 2014 datasets

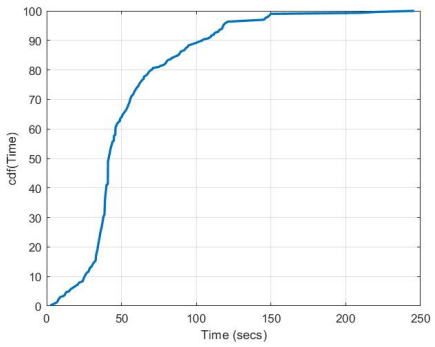
Features	**RMSE**		**MAE**	
	Dev	Test	Dev	Test
AVEC 2017				
Vert	6.24	7.03	4.91	5.84
Diag	6.21	7.14	4.94	5.54
Avg. of Vert & Diag	6.23	7.22	4.86	5.72
AVEC 2014				
Vert	11.58	12.51	9.93	10.36
Diag	11.59	12.60	9.64	10.32
Avg. of Vert & Diag	11.62	12.37	9.93	10.16

for all landmarks located on the eyebrow. Finally, we report that eyelid movement features, crafted to quantify eye openings, are not as effective as predicting depression severity scores as compared to head, mouth, and eyebrow craniofacial movement features.

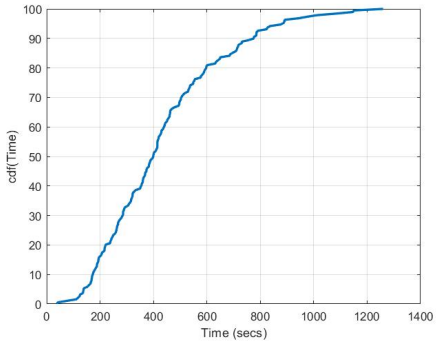
Comparison with published literature

In order to compare our proposed craniofacial movement features, we summarise the contributions from other features in Table 4.7 for AVEC 2017 DSC and Table 4.8 for AVEC 2014 DSC challenges. It is important to mention here that we only list results from the baseline papers of these challenges as well as publications which use features derived, at least in part, from facial landmarks. Our aim is to provide a balanced comparison with craniofactial movement features which only used facial landmarks. It is reminded that we already provide a survey of publications from AVEC 2014, 2016, and 2017 challenges in Section 4.6.

Moreover, it is also pertinent here to state the differences between the two datasets. For example, AVEC 2017 DSC dataset consists of video recordings of structured interview sessions between subjects and a virtual agent. The interview sessions were devised in such a way that subjects were asked questions from the PHQ-8 [21] questionnaire, therefore, the sessions closely followed the protocol for assessment of depression. The AVEC 2014 dataset, meanwhile, consists of video recordings of subjects performing two types of speaking tasks. In the first task, subjects read out a passage from the German fable *North Wind and the Sun*. The second task required subjects to provide answers to questions prompted through a Power Point presentation file. These questions included "*What is your favourite dish?*"; "*What was your best gift, and why?*",



(a) AVEC 2014 dataset



(b) AVEC 2017 dataset

Figure 4.7: Cumulative Density Function of time durations of video recordings in AVEC 2014 and AVEC 2017 datasets

and "*Discuss a sad childhood memory*". Essentially, the sessions from AVEC 2014 DSC do not follow the protocol for depression screening as the AVEC 2017 DSC does. Moreover, we note that the duration of sessions for AVEC 2017 and AVEC 2014 are also different, as illustrated in Figure 4.7. More than half of the recordings in AVEC 2014 dataset have a duration of less than 50 seconds, whereas half of recordings in the AVEC 2017 dataset have a duration of at least 300 seconds. The smaller duration of video recordings for AVEC 2014 dataset means that there is smaller amount of data available to compute craniofacial movement features. Finally, whereas AVEC 2017 dataset included a bespoke recording environment, recordings for AVEC 2014 dataset were made through Skype-like sessions.

In Table 4.7 we summarise the results obtained in the baseline paper for the AVEC 2017 DSC as well as those publications where facial landmarks based features were used. It is important to mention here that since the AVEC 2017 challenge used the same dataset as the AVEC 2016 challenge[1], some researchers

[1]The objectives of these challenges were different: for AVEC 2016 the objective was to

reported results in terms of RMSE/MAE for the AVEC 2016 challenge as well. Therefore, we have also included results from such publications in Table 4.7.

In order to compare our proposed features based facial landmarks, we summarise the contributions from other researchers who have proposed cranio-facial features computed using facial landmarks. To this end, we have summarised results from three AVEC depression challenges.

Ringeval et al. [13] introduce the AVEC 2017, however, since AVEC 2016 and AVEC 2017 challenges use the same dataset, therefore, Ringeval et al. propose the same baseline for visual features as that proposed by Valstar et al. [12] who utilised a combination of facial landmarks based features and appearance features based on LGBP-TOP [81].

For features based on facial landmarks, they computed what they called *geometric features*. These features included a large number of distance and angle measurements between pairs of facial landmarks around the eyes, eyebrows, and the mouth region. They also computed the distance between the median of stable landmarks and each of the 67 other facial landmarks. Thus in total, they compute 316 geometric features and later use functionals based feature aggregation.

We find the proposal from Valstar et al. to compute geometric features for only specific regions of the face particularly interesting, given that prior trend was to compute an exhaustive set of geometric features for all possible pairs of facial landmarks [104, 207]. However, (a) Valstar et al. do not provide details of how they computed features for sub-regions of the face, and (b) they still compute an extensive set of features even if it is not exhaustive.

On the development partition, Valstar et al. achieved an RMSE/MAE = 7.13/5.88, and an RMSE/MAE = 6.97/6.12 on the test partition. However, it is also important to mention here that the efficacy of their proposed *geometric features* cannot be ascertained since they combined these features with LGBP-TOP based appearance features based on LGBP-TOP for predicting depression severity. Nevertheless, we aim to compare the performance of our proposed craniofacial movement features with their work since it forms the baseline for the AVEC 2017 challenge.

Nasir et al. [152] investigate the use of facial landmarks for predicting depression. To this end, they experiment with three feature sets derived from facial landmarks. The first feature set, which they called *facialtracker* is based on taking the average value of each facial landmark over a 10 second temporal window. For the second feature set, which they called geometric feature set, they compute distance and area measures using facial landmarks. The distance features quantify: (a) distance between eyelid and the eyebrow, distance between nose and upper lip, distance between chin and the lower lip, distance between upper and lower lips, and the distance between mouth

classify individuals as depressed or not depressed, whereas for AVEC 2017, the objective was to predict depression severity score

corners. These distance measures are computed for every pair of frames and the average value is taken over a 10 second window. As illustrated in Figure 1 of their paper, a total of 26 features are computed which represent the area encompassed by 68 facial landmarks. The third feature set is based on polyonimal parameterisation of changes in facial landmarks. This involves fitting a second order polynomial function to the value of facial landmarks over a 10 second window. The parameters of the polynomial function are subsequently used as features. Nasir et al. report that amongst the three types of feature sets, the combination of distance and area features provided the best performance (in terms of average F1 score) for the task at hand. This is followed by facialtracker and the polynomial parameterisation featureset. On the development partition, they also report accuracy in terms of RMSE/MAE after combining features from the geometric and facialtracker featureset. Here we note that the feature set achieves an RMSE = 7.86 and an MAE = 6.48. They do not, however, report results for the test partition.

Yang et al. [149], compute geometric features based on distance and angle measurements between pairs of facial landmarks in the region of eye, eyebrows, and mouth. However, they do not provide further information about how distance and angle features are computed. Next, they reduce the dimensionality of their geometric features by applying a PCA such that 99.9% of variance is retained. Finally, they take the average value of the PCA-geometric features over the entire duration of recording in order to create a global representation of video. Yang et al. report the performance of the computed features for the development partition for gender dependent models. As summarised in Table 4.7, the model for females achieves an RMSE/MAE = 6.39/5.10, whereas the model for males achieves an RMSE/MAE = 6.99/5.75. Similarly, Sun et al. [153] compute PCA-geometric features (without providing details), and achieve an RMSE/MAE = 7.06/5.77 on the development partition.

In Table 4.7, we provide performance comparison of our proposed craniofacial movement features with the works Valstar et al. [12], Nasir et al. [152], Yang et al. [149], and Sun et al. [153]. Here, we observe that our proposed craniofacial movement features achieve best performance in terms of RMSE on the development partition. Head movement, Mouth movement, and Eyebrow movement features also beat the challenge baseline on the test partition in terms of both RMSE and MAE. Amongst these results, the most interesting observation was that craniofacial movement facial features were able to beat the combination of geometric and appearance features proposed by Valstar et al.

Finally, given that our features are also interpretable, that is, one can explicitly link the contribution of various aspects of facial movement to depression severity score, we believe that these features can provide meaningful feedback to clinicians for diagnosis of depression.

In Table 4.8 we summarise the results obtained in the baseline paper for the AVEC 2014 DSC as well as those publications where facial landmarks based

Table 4.7: Performance comparison of proposed craniofacial movement features for AVEC 2017

Features	**RMSE**		**MAE**	
	Dev	Test	Dev	Test
Valstar et al. [12] (*baseline*)	7.13	6.97	5.88	6.12
Nasir et al. [152]	7.86	–	6.48	–
Yang et al. [149] (female)	6.39	–	5.10	–
Yang et al. [149] (male)	6.99	–	5.75	–
Sun et al. [153]	7.06	–	5.77	–
Head Movement	6.05	6.27	4.96	5.32
Mouth Movement	6.05	6.20	4.98	4.94
Eyelid Movement	6.24	7.03	4.91	5.84
Eyebrow Movement	6.27	6.29	5.26	5.13

features were used. We start with the work of Valstar et al. [11] who introduced the AVEC Depression Recognition Challenge and provide baseline results for the visual modality. They utilise LGBP-TOP features [81] to quantify facial movement. The fundamental idea behind this approach is that facial muscle movement manifests as changes in skin texture with appearance of facial furrows and wrinkles, which can be quantified through the use of texture-level features. On the development partition, Valstar et al. achieve an RMSE = 9.26 (they do not report MAE for the development partition), meanwhile on the test partition, these features provide an RMSE/MAE = 10.86/8.86. It is worth mentioning here that since Valstar et al. did not use features derived from facial landmarks therefore a direct comparison between our work and theirs is not possible. Nevertheless, it is still important to compare our work with theirs since their work is the challenge baseline.

The only publication from AVEC 2014 challenge which used facial landmarks based features is the work of Gupta et al. [223]. They hypothesised that head and overall facial movement carry important information about affective state and depression level. They quantify this movement using LBP [291], LBP-TOP [81], and optical-flow-based motion vectors [209] between a pair of consecutive frames at key points computed using a corner detection algorithm [292]. They also computed pairwise Euclidean distances between each of the 66 facial landmarks and a stable point between the eyes. As per results reported in their paper, Gupta et al. achieved an RMSE of 9.00 on the development partition, which is better than any of our proposed craniofacial movement features as shown in Table 4.8.

A caveat to their result is the fact that they used features based on facial landmarks alongside multiple appearance based visual features such as LBP, LBP-TOP, and optical flow, thus combining four types of features through

Table 4.8: Performance comparison of proposed craniofacial movement features for AVEC 2014

Features	RMSE		MAE	
	Dev	Test	Dev	Test
Valstar et al. [11] (*baseline*)	9.26	10.86	–	8.86
Gupta et al. [223]	9.00	–	–	–
Head Movement	10.71	10.96	9.22	9.04
Mouth Movement	10.12	11.19	8.50	9.37
Eyelid	11.58	12.51	9.93	10.36
Eyebrow	11.11	11.80	9.28	9.71

feature-level fusion. In addition to this, Gupta et al. also used a two-stage greedy feature selection algorithm to optimise RMSE for the development partition. Thus one cannot separate the contribution of appearance based visual features from the contribution of facial landmarks based features.

In Table 4.8, we provide performance comparison of our proposed craniofacial movement features with the works Valstar et al. [11] and Gupta et al. [223]. Here we note that results of our proposed features do not hold particularly well i.e. our best result i.e. mouth movement features, trails both Valstar et al. and Gupta et al. Given the results achieved on the AVEC 2017 dataset, where these features demonstrated strong performance, the results achieved for AVEC 2014 dataset are disappointing.

Nevertheless, as discussed earlier in this section, the AVEC 2017 dataset follows the clinician-patient interaction protocol much more closely as compared to the AVEC 2014 dataset. This could be one of the reasons why our features perform better for the AVEC 2017 dataset. Nevertheless, our proposed features offer the unique advantage of interpretability, that is, one can explicitly link the contribution of various aspects of face based motor activity to depression severity score.

4.8.2 Turbulence Features for Audio Modality

We compute turbulence features over five different time-scales, and use five descriptive statistics to summarise their values, as detailed in Section 4.7.1. Therefore, the resultant feature has 25 dimensions.

In order to build a model for prediction of depression severity, we use PLSR [172]. We vary the number of components between 4 and 8, and optimise for the objective of achieving the smallest RMSE on the development partition. There is, however, an interesting question which exists with the use of these features i.e. should unvoiced frames be removed from the features or should they be retained. On one hand it makes sense to remove them because

Table 4.9: Performance of turbulence features while keeping unvoiced frames as zero

Feat	**Train**			**Dev**			**Comp**
	MAE	*RMSE*	*Corr*	*MAE*	*RMSE*	*Corr*	
F0	4.23	5.20	0.29	4.81	5.95	0.43	4
NAQ	4.14	5.06	0.37	5.03	6.16	0.37	8
QOQ	4.24	5.23	0.27	6.2	8.29	0.04	4
H1H2	4.11	5.16	0.32	5.27	6.47	0.13	8
PSP	4.39	5.25	0.26	6.52	10.40	0.04	4
MDQ	4.04	5.05	0.37	5.21	6.43	0.21	4
PS	4.23	5.12	0.34	5.19	6.33	0.30	5
Rd	4.41	5.34	0.19	5.33	6.43	0.20	7

Table 4.10: Performance of turbulence features after removing unvoiced frames

Feat	**Train**			**Dev**			**Comp**
	MAE	*RMSE*	*Corr*	*MAE*	*RMSE*	*Corr*	
F0	4.40	5.21	0.29	5.58	6.75	0.01	4
NAQ	4.31	5.34	0.19	5.33	6.40	0.27	7
QOQ	3.74	4.79	0.47	5.66	6.73	0.02	7
H1H2	4.07	5.14	0.33	5.38	6.54	0.12	5
PSP	4.00	4.88	0.44	6.22	8.22	0.10	7
MDQ	4.22	5.17	0.31	5.45	6.49	0.20	7
PS	3.93	4.79	0.48	5.55	6.95	0.10	4
Rd	4.06	4.93	0.42	5.49	6.64	0.03	4

they will contain information not related to speech, but on the other hand, keeping unvoiced frames may provide information about the rhythm of an individual's speech.

We empirically test two approaches, with results summarised in Tables 4.9 and 4.10. In Table 4.9, instead of removing unvoiced frames, we simply change their value to zero, meanwhile in Table 4.10, we remove those frames altogether. An inspection of these two tables suggests that it is almost always beneficial to retain unvoiced frames if their value is changed to zero. The biggest beneficiary are features derived from F0, which beats the best RMSE and MAE scores for the development partition, as given in the baseline paper.

4.8.3 Fisher Vector encoding of Spectral LLDs

In this section, we discuss experiments on using spectral modelling of speech of depressed individuals using Fisher Vector encoding, for AVEC 2016 DCC

and AVEC 2017 DSC datasets.

Classification Task: The AVEC 2016 DCC

For the classification task, we report on our approach for the AVEC 2016 depression severity prediction sub-challenge [12]. The objective of this challenge was to classify between individuals who had depression and those who did not. The results were measured using the F1 score. It must be mentioned here that these experiments were performed using training and development subsets since these experiments were performed after the challenge was formally closed and the organisers were no longer providing results on the test. Nevertheless, as discussed below, we use create multiple partitions of data available to test our proposed approach.

We start by combining data from the training and the development sets and create 10 randomly generated partitions (so-called sub-train and sub-test), while maintaining 80/20 ratio of non-depressed/depressed classes. We also retain the same number of samples in the two sub-sets as provided by the organisers i.e. 107/35 for the training and development. It was ensured that data from sub-training partition does not mix with the data from the sub-test partition. Next, optimal values of hyper-parameters of WELM are selected using grid-search over the set $\{50, 100, \ldots, 1000\}$ for L and $\{2^{-3}, 2^{-2}, \ldots, 2^{9}\}$ for c and the number of clusters for the GMM used for FV encoding were selected as $\{4, 8, 16, 32, 64\}$, for each partition, with the objective of maximising the mean F1 score. We report classification results in Table 4.11 of FV encoding of various spectral features as the median value of their score on the 10 partitions.

We note that the best result is achieved by FV encoding of PLP features when 64 cluster GMM is used to model the background, although even with 4 clusters, PLP features once FV encoded provide a better performance than the AVEC 2016 DCC baseline, as shown in Table 4.12. FV encoding of MFCCs, another popular spectral representation of speech, is also able to improve on the baseline, however, its performance lags behind that of PLPs.

Surprisingly, formants failed to yield satisfactory results. One expected better results since formant frequencies have been described as correlates of vocal tract [130], which could have been more discriminatory. We suspect that their performance may have been compromised due to free speech nature of the dataset. Meanwhile, features based on MCEPs, HMPDMs, and HMPDDs, provided as part of the dataset, are also unable to reach the baseline performance.

Finally, when compared to results from other publications for the AVEC 2016 DCC, our results are comparable to the best performance on the development partition. In fact, if one is to consider the maximum score achieved on the 10 partitions, then we achieve the best result on the development partition. While the results are exceptionally good, one must appreciate the fact that machine learning algorithms have a tendency to overfit when the size of the

Table 4.11: Classification results on the development partition with FV encoding and WELM

Features	GMM 4	GMM 8	GMM 16	GMM 32	GMM 64
MFCCs	0.704	0.761	0.800	0.788	0.807
PLPs	0.791	0.842	0.857	0.885	**0.906**
Formants	0.551	0.622	0.630	0.628	0.636
MCEPs	0.589	0.619	0.661	0.649	0.687
HMPDMs	0.636	0.627	0.621	0.642	0.617
HMPDDs	0.596	0.628	0.599	0.608	0.612

Table 4.12: Comparison of results as mean F1 for publications from AVEC 2016 DCC

Publication	Development	Test
(baseline) [12]	0.698	0.720
Ma et al. [305]	0.610	—
Pampouchidou et al. [150]	0.730	0.665
Nasir et al. [152]	0.760	—
Huang et al. [155]	0.765	0.550
Williamson et al. [130]	0.810	0.700
Our work *(median)*	**0.906**	—
Yang et al. [149]	0.910	0.724
Our work *(max.)*	**0.957**	—

dataset is small, which is typically the case in most social signal processing domains. Nevertheless, the efficacy of our proposed method is evident through these results.

Regression Task: The AVEC 2017 DSC

For the regression task, we report on our approach for the AVEC 2017 depression severity prediction sub-challenge [13]. The objective in this challenge was to predict the PHQ scores of individuals on the test partition. The performance was measures in terms of the MAE and RMSE.

While we computed a single Fisher Vector for spectral LLDs for the classification task discussed in previous section, we proposed a multiscale FV encoding of spectral LLDs, and use pooled functionals from these Fisher Vectors as features for predicting depression severity.

The background modelling part is similar, and it is only the Fisher Vector encoding part which is different for the multi-scale approach. In order to achieve this, we take spectral LLDs windows of $\{0.5, 5, 10\}$ seconds, with overlap of

Table 4.13: Summary of results for various pooling methods for multi-scale FV encoding.

Pool .	**Train**			**Dev**			**Res.**	**GMM**
	MAE	*RMSE*	*Corr*	*MAE*	*RMSE*	*Corr*		
Mean	3.05	3.38	0.90	5.53	6.50	0.43	3	16
Max	4.81	5.66	0.60	5.67	6.52	0.33	3	16
Med.	4.81	5.66	0.54	5.65	6.52	0.32	2	24
CF	4.81	5.66	0.71	5.66	6.52	0.34	1	32
Range	4.78	5.47	0.59	5.60	6.42	0.37	3	16
Var	4.81	5.66	0.70	5.65	6.52	0.41	2	16

$\{0.2, 3, 5\}$ seconds, respectively. The FVs computed over each of these time scales are pooled into a single FV by applying the following descriptive statistics element-wise: mean, max, median, variance, crest factor, and range.

We use SVR, to build models based on Fisher vectors which can predict PHQ scores for individuals. We train SVRs with an RBF kernel. We utilise Matlab wrappers for ε-SVR available as part of the libSVM [169] toolkit. The cost parameter C is searched between $2^{\{-10:10\}}$, *Epsilon* between $2^{\{-5:5\}}$ and the width of the RBF kernel *Gamma* between $2^{\{-16:4\}}$, with a step of 2. Amongst these, we select the parameters which yield the largest absolute Pearson correlation values on the development partition.

In Table 4.13, we provide a summary of the best performing pooling methods on the training and development sets in terms of RMSE and MAE as well as the absolute Pearson correlation coefficients, whilst selecting a cut-off p-value of 0.05. We note that all pooling methods are able to achieve virtually similar performance in terms of MAE and RMSE, when one has options to choose any particular temporal resolution for FV encoding along with the number of clusters for the GMM.

There are, however, subtle differences. Firstly, we note that the absolute Pearson correlation value of 0.43 on the Development partition is achieved through Mean pooling of FVs, when using *Res* 3 i.e. a window of 10 seconds and using a 24 cluster GMM. Mean pooling also has the smallest MAE and RMSE on the development partition compared to other pooling methods. Another important observation is that most pooling methods perform well at *Res* 3 i.e. a temporal resolution of 10 seconds, while the crest factor stands out as the only pooling method which works best at a temporal resolution of 500 ms. We believe that this is due to the nature of the crest factor, which essentially measures turbulence, and at larger time-scale, micro-level description is not as fruitful for the task of predicting labels.

While the objective of the AVEC 2017 DSC was to achieve the smallest possible RMSE, we believe that there may be cases where one may want to

Table 4.14: Trade-off between minimising RMSE and maximising correlation.

Pooling	**Train**			**Dev**		
	MAE	*RMSE*	*Corr*	*MAE*	*RMSE*	*Corr*
Max (best RMSE)	4.81	5.66	0.60	5.67	6.52	0.33
Max (best corr)	4.45	5.62	0.89	5.45	6.93	0.42
Mean (best RMSE)	3.05	3.38	0.90	5.53	6.50	0.43
Mean (best corr)	3.67	3.87	0.92	5.50	6.55	0.46

measure depression severity using a parameter which may not match PHQ scores in terms of its dynamic range (therefore have poor RMSE), but closely matches the PHQ labels through correlation. For example, consider Table 4.14, where we summarise possible trade-offs between the choice of smaller RMSE or a higher absolute Pearson correlation.

4.9 Summary

In this chapter, we introduced three new approaches for the task of automated depression screening from audio-visual recordings, namely; turbulence features, craniofacial movement features, and Fisher Vector encoding of spectral LLDs. The efficacy of these methods was demonstrated using datasets from 2014, 2016, and 2017 editions of the AVEC challenges on depression screening.

To conclude, we summarise the key achievements of our work as follows:

- We surmised that psychomotor changes due to depression lead to uniqueness in an individual's speech pattern which manifest as sudden and erratic changes in speech feature contours. To this end, we proposed a novel set of temporal features, which we called turbulence features, to quantify fluctuations in the feature contours of speech features.

  The efficacy of turbulence features was demonstrated as part of our solution for the AVEC 2017 DSC [13], where we stood 6th overall in the competition, beating the challenge baseline [34]. Amongst various voice quality and prosody features which were part of our investigation, we found turbulence features computed for pitch feature contour to be most useful for the task of automated depression screening.

- We detailed a methodology to quantify specific craniofacial movements, which we hypothesised could be indicative of psychomotor retardation and hence depression.

  The efficacy of these features was tested in terms of the value of Pearson's correlation coefficient with respect to depression severity. We used three

sets of recordings from two publicly available datasets from AVEC 2014 DSC [12] and AVEC 2017 DSC [13].

The results demonstrated the efficacy of our proposed craniofacial movement features. Moreover, given that these features are inspired by knowledge of psychomotor retardation from the DSM 5 manual [18], we believe that interpretability of these features will provide meaningful feedback to clinicians for diagnosis of depression.

- We hypothesised that individuals with depression have unique characteristics to their speech spectra. To this end, we introduced Fisher vector encoding of spectral LLDs for quantifying abnormalities within speech spectra of individuals with depression.

  Initially, we demonstrated the efficacy of our proposed approach for the AVEC 2016 DCC dataset [12], where the objective was to identify individuals with and without depression [37]. Later, we extended the idea by adding temporally-piecewise aggregation of Fisher vectors as part of our solution to the AVEC 2017 DSC [34]. We beat the challenge baseline whilst using this method.

- We note that datasets released as part of AVEC sub-challenges on automated depression screening (2014, 2016, and 2017 editions) have all used accumulated scores from various depression measurement instruments as labels for audio/visual recordings. The AVEC 2014 DSC dataset [11] used BDI-II [38], whereas the AVEC 2016 DCC dataset [12] and AVEC 2017 DSC dataset [13] used PHQ-8 [21]. While these scores provide a way to quantify depression severity in absence of physical tests for depression [28], there are inherent limitations of using these labels.

  For example, the BDI-II [38] asks patients about feelings of satisfaction, disappointment, and guilt, as well as questions about their weight, appetite, and sex-life. It is obvious that information pertaining to these questions cannot be extracted from audio/visual recordings unless patients are explicitly recorded whilst answering these questions. In that case, one would use speech-to-text conversion and follow it up with natural language processing to learn the answers to these questions.

  This means that under the current set up of datasets, automated screening methods based on audio/visual modalities will continue to have sub-par performance relative to ground truth labels. We argue that it may be worth investing time to devise depression measurement instruments (questionnaires) which specially cater for development of automated screening methods. This would, of course, require significant collaboration between researchers from psychology and SSP/AC.

- In light of our discussion in this chapter and Chapter 3, we believe it is important to emphasise here that while significant inroads have been

made for the task of automated screening of depression, this task is still very much a work in progress.

The most outstanding issue remains the lack of publicly available datasets, which is further exasperated by potentially noisy labels from self-administered depression assessment instruments. The many confounding factors such as gender, age, and the nature of speaking tasks means that research for the development of automated methods for screening of depression is likely to continue for at least the near future before these methods are deemed ready for clinical usage.

Chapter 5

Automated Screening for Bipolar Disorder

5.1 Introduction

Bipolar disorder is a mental disorder, which, according to the World Health Organisation (WHO), affects more than 60 million individuals worldwide — making it amongst the top-ten mental disorders for adults worldwide [325].

Individuals with bipolar disorder suffer from episodes of depression and mania, which are separated by periods of normal mood. Manic episodes typically include irritable mood, hyper-activity, loud speech, inflated self-esteem and a decreased need for sleep [326, 327]. While bipolar disorder is a life-long illness, early diagnosis and subsequent treatment can favourably impact the quality of life for individuals with this mental disorder. This is where automated screening methods can help.

Conventional methods for screening of bipolar disorder are subjective in nature i.e. based on self-assessment or clinician-assessment based questionnaire. Based on our discussion in Section 1.4, our motivation is to propose automated screening methods for this disorder. Accordingly, we propose various methods for automated screening of bipolar disorder from audio/visual modalities. This is our solution for the Bipolar Disorder sub-challenge (BDS) which was part of the Audio/Visual Emotion recognition Challenge (AVEC) 2018, co-located with ACM Multimedia Conference.

The AVEC 2018 BDS is the very first of its kind within the scope of mental health analysis where the objective is to predict severity of mania for individuals with bipolar disorder using audio/visual recordings of structured interviews. The reader is reminded that previous editions of AVEC challenges focussed on automated screening of depression [10–13].

The layout of this chapter is as follows: we start with statements of novelty and contributions through our work on development of automated screening of bipolar disorder. We follow this by describing traits of individuals with bipolar

disorder according to DSM-5 manual [18]. We also briefly describe states of bipolar disorder as per the Young Mania Rating Scale (YMRS) [22]. This provides the foundation for our proposed automated screening methods. Next, we provide a survey of research literature published for automated screening of bipolar disorder from audio/visual modalities. While limited research literature exists, we were able to identify some features from audio/visual modalities which were previously deemed useful by others. Next, we discuss our proposed methods in detail and provide experimental analysis. We also discuss our submissions for the test partition of the AVEC 2018 Bipolar Disorder sub-challenge. Finally, we provide a conclusion based on insights from our work.

5.2 Novelty and Contributions

Our proposed approaches for automated screening of bipolar disorder from audio/visual modalities is inherently novel for this task since the AVEC 2018 Bipolar Disorder sub-challenge (BDS) [33] provides researchers for the very first time a dataset which contains multi-modal recordings of individuals with bipolar disorder based on structured interviews. The contributions of our work can be listed as follows:

- We propose *turbulence features* to capture sudden and erratic changes in feature contours, and demonstrate its efficacy for the task at hand. The reader is reminded that we had initially proposed these features for the task of predicting depression severity at AVEC 2017 (see Section 4.7.1 for details).

- We introduce Fisher Vector encoding of Computational Paralinguistics Challenge (ComParE) low level descriptors and demonstrate that these features are viable for predicting the severity of mania. In fact, we show that these features perform significantly better than ComParE functionals [64] for the AVEC 2018 BDS.

  The reader is referred to Section 5.7.3 for details of Fisher Vector features based on ComParE LLDs and to Section 5.7.3 for experimental results.

- We introduce the Greedy Ensembles of Weighted Extreme Learning Machines (GEWELMs) classifier, for the task of classifying individuals with bipolar disorder into states of remission, hypo-mania, and mania.).

  The reader is referred to Section 5.6.4 for details of GEWELMS and to Section 5.7 for experimental results.

- The best result on the test partition i.e. UAR = 57.41% is achieved whilst using the proposed turbulence features, computed for features pertaining to facial movement and associated emotions. This result

exactly matches the baseline of the AVEC 2018 Bipolar Disorder sub-challenge. It also matches the state-of-art for the sub-challenge, as discussed in Section 5.7.6.

5.3 Bipolar Disorder as per the DSM-5 and the YMRS

Bipolar disorder is a complex mental disorder which modulates the overall behaviour of an individual — in particular, their mood. According to the DSM-5 manual [18], bipolar disorder is characterised by cyclic periods of depression followed by episodes of mania.

Mania is a state of elevated mood, arousal, affect, and energy levels, and its symptoms are generally opposite to those reported for state of depression [18]. For example, an individual in a manic state has higher-than-normal energy, elevated mood, decreased requirement for rest/sleep, and is more socially active. In depressive state, meanwhile, an individual typically goes into social withdrawal, has decreased arousal, and is generally more passive in their behaviour and social activities.

The existence of depression can be gauged using the DSM-5 guidelines for major depressive disorder, which we discussed previously in Section 4.3 as part of our work on automated screening of depression in Chapter 4. The existence of mania and its severity is commonly quantified using the Young Mania Rating Scale (YMRS) [22, 33, 39].

YMRS is a MCQ style questionnaire which was developed by Young et al. [22] to assess manic symptoms. It is similar to the Hamilton Rating Scale for Depression (HAM-D) [20] questionnaire which is used for assessment of depression severity. The YMRS contains 11 items in total. Amongst these, four items are graded on a scale between 0–8 (querying irritability, speech, thought content, and disruptive/aggressive behaviour), whereas the remaining seven items are graded on a scale between 0–4. The YMRS can either be completed by the psychiatrist (psychiatrist-administered) or the individual affected with bipolar disorder (self-administered). For the dataset provided as part of the AVEC 2018 BDS, YMRS was completed for subjects by trained psychiatrists [33]. The reader is directed to an online version of YMRS [23] for further insights about the questionnaire.

The severity of bipolar disorder is typically measured in terms of three sub-states, which are remission, hypo-mania, and mania [33]. Remission is the lowest severity of bipolar disorder, and it represents suppression of symptoms through medication. In terms of the YMRS, an individual with YMRS $\leq$ 7 is considered to be in the remission state of bipolar disorder [33, 39]. An individual with YMRS $>$ 7 is considered to be in a state of mania. Here, one is abnormally energised, both mentally and physically. When the YMRS score is between 8 and 20, an individual is said to be in the state of hypo-

mania, whereas a score YMRS $>$ 20 means that an individual is the state of mania. Simply put, symptoms of mania are much more intense than those of hypo-mania.

5.4 Literature Survey

As mentioned earlier, the AVEC 2018 BDS is the very first of its kind in the field of social signal processing to provide a publicly available dataset with audio-visual recordings of structured interviews of individuals with varying severity of bipolar disorder. Given this fact, one finds limited research work in this field so far. We foresee numerous publications based on this dataset in near future. However, we do find the following publications relevant for the task of automated screening of bipolar disorder.

Guidi et al. [239] investigated voice quality of individuals with bipolar disorder using *long-term spectral average* features. They report statistically significant differences over certain intervals of speech spectra for individuals with bipolar disorder, thereby suggesting that speech spectra can be used to screen for bipolar disorder. In a more recent publication, Guidi et al. [183] undertook spectral analysis of pitch (F0) contours extracted from audio recordings of bipolar patients and healthy control subjects as both read emotion inducing texts. They report that analysing pitch contours was an effective method to identify individuals with bipolar disorder.

Zhang et al. [184] investigated the efficacy of pitch, formant frequencies, and spectral features in terms of linear prediction coefficients (LPC) [328] for the task of distinguishing between individuals with bipolar disorder and healthy controls. They report that the first, second, and fourth formant frequencies, along with LPC features performed well for the task at hand.

Meanwhile, Maxhuni et al. [226] investigated classification of episodes of bipolar disorder from voice and motor activity using data collected from smart phones. For this, they collected audio from voice calls and accelerometer data along with self-assessment data from five patients over a period of 12 weeks. They demonstrate that it is possible to identify manic episodes with an accuracy of 85% for their dataset.

Ciftci et al. [33] introduce the Turkish Audio-Visual Bipolar Disorder Corpus, a subset of which forms the dataset used for AVEC 2018 BDS. The dataset is annotated for sub-state of bipolar disorder as well as scores from the Young Mania Rating Scale (YMRS). In order to capture information from the visual modality, they computed geometric features from facial landmarks and texture-based features from a fine-tuned Deep Convolutional Neural Network. For audio modality, they used Interspeech 2010 paralinguistic challenge features (IS10) acoustic feature [329]. Partial least squares regression [103] and extreme learning machines [123] for classification.

5.5 Dataset

The dataset provided as part of the AVEC 2018 Bipolar Disorder sub-challenge (henceforth, we shall refer to this dataset as the AVEC 2018 BDS dataset) is a subset of the Turkish Audio-Visual Bipolar Disorder Corpus [33], which was published at the Asian Conference on Affective Computing and Intelligent Interaction 2018 (ACII Asia 2018) [330].

Participants of the AVEC 2018 BDS have to classify patients suffering from bipolar disorder into three classes i.e. states of remission, hypo-mania, and mania from audio-visual recordings of structured interviews. The AVEC 2018 BDS dataset contains 218 audio-visual recordings in total. Amongst these, 104 recordings form the training partition, 60 form the development partition, and the remaining 54 form the test partition. We note that while subjects have multiple recordings within each partition, subjects are not repeated across partitions.

Each recording contains multiple sessions in which subjects are asked to perform various speaking tasks. These tasks range from describing happy and sad memories, counting numbers up to thirty, and explaining emotion eliciting pictures [33, 39]. The dataset was annotated for bipolar disorder state on the basis of Young Mania Rating Scale (YMRS) by trained psychiatrists. The YMRS score of each subject was subsequently quantised into three sub-states of bipolar disorder using the following set of rules [33]:

1. Remission: YMRS ≤ 7
2. Hypo-mania: $7 <$ YMRS ≤ 20
3. Mania: YMRS > 20

It is also important to mention here that no baseline paper was released for AVEC 2018 BDS while the challenge was ongoing but organisers did release a preprint of Ciftci et al. [33] to facilitate participants of the challenge by providing them fundamental information about the dataset. One expects the baseline paper to be released soon as [39].

5.6 Methodology

The objective of the AVEC 2018 BDS is to predict the severity of mania for individuals with bipolar disorder. Given that labels are based on YMRS scores (see Section 5.5 for details), we first identify key behavioural characteristics of individuals with mania as per the YMRS. This enables us to craft features which can probe for the existence of these characteristics, as opposed to brute-force methods which aim to learn the dataset without using background knowledge. Amongst the 11 questions in the YMRS questionnaire, we find the following characteristics to be most relevant: elevated mood, increased motor activity,

irritability, speech (in terms of rate and amount), and disruptive-aggressive behaviour.

In order to probe these behavioural characteristics, we propose to use the following methods:

Turbulence features: To quantify erratic changes in feature contours of pitch, formants, eye-gaze, headpose, and facial action units.

Fisher Vector encoding of ComParE LLDs: To quantify prosodic, voice quality, and spectral characteristics of speech.

openSmile standard feature sets: We use four standard feature sets from the openSmile toolkit [30] to quantify characteristics of speech which have been proven to change with respect to emotional and paralinguistic changes in speech. Our aim is to not only determine the efficacy of these feature sets for the task of predicting severity of mania, but also to establish a baseline to compare turbulence features and Fisher Vector features.

5.6.1 Turbulence Features

We hypothesise that feature contours for individuals under remission, hypomania, and mania states of bipolar disorder are essentially different, and if quantified appropriately may provide an insight into the severity of mania. We propose to use turbulence features for this purpose. It is reminded that we had first proposed turbulence features for the task of predicting depression severity from audio modality, as discussed earlier in Section 4.7.1.

Here, we extend the idea by computing *turbulence features* for features from both audio and visual modalities. We start by selecting features from audio/visual modalities which, we surmise, are representative of behavioural characteristics of mania:

Visual Modality

We compute features for visual modality using the OpenFace toolkit [32]. OpenFace provides a large set of features which include facial landmarks, evidence of activation for facial action units, eye-gaze, head-pose, and histogram of oriented gradient (HOG) [78] features for the face region. Amongst these, the following subset of features is selected, for which we subsequently compute turbulence features.

- Eye-gaze and Headpose features are used to quantify motor activity in terms of eye- and head-movement, respectively.
- AU6 and AU12 features represent the cheek raiser and the lip corner puller, respectively. These action units are activated for facial expressions

which capture emotions of happiness as per the facial action coding system (FACS) [2, 40]. We surmise that these features can be used to quantify elevated mood, in addition to activation of facial muscles.

- AU2 and AU5 features represent the outer brow raiser and upper lid raiser, respectively. As per FACS, these action units are activated during facial expressions of high arousal such as anger, surprise and fear. We surmise that these features can be used to quantify elevated mood and irritability, in addition to activation of facial muscles.

- AU20 feature represents the lip stretcher. According to FACS, it is activated for facial expressions representing emotions of fear. We surmise that evidence of AU20 can be used to quantify mood changes of each subject.

- Finally, the AU45 represents blinking. We surmise, based on [331], that evidence of AU5 can be used to gauge sleepiness of each subject.

Audio Modality

For the audio modality, we use feature contours of pitch and the first five formant frequencies, both of these are computed using the COVAREP toolkit [62].

Pitch quantifies the rate of vibration of the vocal folds, and perceptually represents the melodic contour of speech. We have already demonstrated that turbulence features computed for pitch are useful for predicting depression severity as part of our solution for the AVEC 2017 DSC (see Section 4.8.2, and [34]). Here, we aim to test the efficacy of these features for predicting severity of mania.

Formant frequencies are harmonics of the pitch in the magnitude spectra of speech, and were found to be useful for screening of bipolar disorder by Zhang et al. [184]. Moreover, the first two formants are also correlated with horizontal and vertical tongue movement as per Cohen et al. [277], thus these feature can also capture motor activity.

We make one further change in the method of computing turbulence features for the task of screening for bipolar disorder, compared to our previous work on screening for depression. Given that the duration of audio/visual recordings is shorter for AVEC 2018 BDS dataset ($\sim$ 10 seconds) as compared to the AVEC 2017 DSC dataset ($\sim$ 5 minutes), we use comparatively shorter duration temporal windows with the AVEC 2018 BDS for multi-scale computation of turbulence features. That is, we use a set of temporal windows of lengths $\{0.5, 1.0, 2.0\}$ seconds, with an overlap of $\{0.2, 0.4, 0.8\}$ seconds, respectively.

We demonstrate the efficacy of our proposed turbulence features via experimental analysis in Section 5.7, in particular Table 5.3 and Table 5.4.

5.6.2 Fisher Vector encoding of ComParE LLDs

We had earlier demonstrated the efficacy of Fisher Vector encoding of spectral features for the task of screening for individuals with and without depression, as well as their depression severity from their speech alone (see Section 4.8.3 and [34, 37]). Here, we propose to encode low level descriptors (LLDs) from the Computational Paralinguistics Challenge (ComParE) feature set as Fisher Vectors and use them for the task of predicting severity of mania.

Our motivation for using ComParE LLDs is based on the fact that these LLDs contain features which represent information pertaining to speech prosody, voice quality, and spectral features [30]. Therefore, intuitively, by using all these LLDs, Fisher Vector features should represent even more detailed information about the characteristics of speech for individuals with bipolar disorder, compared to spectral features alone.

ComParE features are well known in the field of social signal processing and were first introduced by Schuller et al. in [64] as the baseline feature set for 2013 edition of the Interspeech Computational Paralinguistics Challenge. While the original feature set used functionals for feature aggregation, we opt to use Fisher Vector encoding for feature aggregation, and demonstrate that our proposed method achieves much better performance in terms of classification accuracy (see Section 5.7) compared to ComParE functionals.

We computed ComParE LLDs using the openSmile toolkit [30] with `ComParE_2016.conf` configuration file and a setting to save LLDs instead of functionals. The list of features which are part of the 65-dimensional ComParE LLDs is provided in Table 5.1 as a reference for the reader, based on descriptions provided in [30, 64, 332].

5.6.3 openSmile Acoustic Feature Sets

The bipolar disorder sub-challenge has for the first time introduced a publicly available (albeit restricted) dataset for developing methods for screening of bipolar disorder. We take this opportunity to perform experiments with four standard features from the openSmile toolkit [30] which are popular in the social signal processing community. These feature sets include: (a) prosody feature set computed using `prosodyShsViterbiLoudness.conf`, (b) Interspeech 2010 Paralinguistics (IS10-Paraling.) feature set [329] computed using `IS10_paraling.conf`, (c) Interspeech 2013 ComParE feature set (ComParE functionals) [64] computed using `ComParE_2016.conf`, and (d) the eGeMAPS [58] feature set computed using `eGeMAPSv01a.conf`.

As their names suggest, the Prosody and IS10-Paralinguistics feature sets are optimised for capturing information relevant to speech prosody and paralinguistic activity, respectively. Meanwhile, ComParE feature set is a 6,373-dimensional feature set, often called a brute-force feature set due to its size and the fact that it was designed as a general purpose feature set to capture

Table 5.1: List of features within 65-dimensional ComParE LLDs

Feature Group	Feature Names	Dimension(s)
Prosody	Pitch	1
	Loudness	1
	Sum of RASTA-style filtered auditory spec.	1
	RMS energy	1
	Zero Crossing Rate (ZCR)	1
Voice Quality	Harmonics to Noise Ratio (HNR)	1
	Jitter	2
	Shimmer	1
	Voicing Probability	1
Spectral	Mel Frequency Cepstral Coeff. (MFCCs)	14
	RASTA-style auditory spectrum	26
	Spectral Energy (250-650 Hz and 1-4 KHz)	2
	Spectral Roll-off point (0.25, 0.5, 0.75, and 0.9)	4
	Spectral Centroid and Variance	2
	Spectral Flux, Entropy, and Slope	3
	Spectral Skewness and Kurtosis	2
	Psychoacoustic Sharpness and Harmonicity	2

paralinguistic activity from speech. It has certainly lived up to its reputation of brute-force feature set, producing excellent results for Interspeech 2017 and 2018 baselines [146, 148], surpassing even deep learning based approaches. However, brute-force is not always desirable (see Chapter 6). Eyben et al. [58] proposed an 88-dimensional minimalistic audio feature set called the eGeMAPS as an alternate to the ComParE feature set. They also demonstrated the efficacy of eGeMAPS on various datasets for the task of emotion recognition from speech.

Thus, by reporting accuracy achieved using each these feature sets, we aim to highlight the role played by features relevant for prosody, paralinguistic activity, and emotions toward classification of individuals with bipolar disorder.

5.6.4 Classification

We used three fundamental types of classifiers for classifying individuals into states of remission, hypo-mania, and mania. These classifiers include SVM1 [116] classifiers with a linear kernel, logistic regression, and Greedy Ensembles of Weighted Extreme Learning Machines (GEWELMs).

SVM and logisitic regression based classification was performed with the aid of the LIBLINEAR toolkit [171], which is a popular open-source toolkit. We were motivated to utilise this toolkit since baseline experiments were performed whilst using it [39]. Here, a grid search was performed between

all seven possible solvers provided by `LIBLINEAR toolkit` [171]. The cost function parameters were optimised using grid-search with $C = \{1 \times 10^{-5}, 2 \times 10^{-5}, 5 \times 10^{-5}, 1 \times 10^{-4}, 2 \times 10^{-4}, 5 \times 10^{-4}, 1 \times 10^{-3}, 2 \times 10^{-3}, 5 \times 10^{-3}, 1 \times 10^{-2}, 2 \times 10^{-2}, 5 \times 10^{-2}, 1 \times 10^{-1}, 2 \times 10^{-1}, 5 \times 10^{-1}, 1 \times 10^{0}\}$, with the objective of maximising accuracy on the development partition.

In the following section we discuss the GEWELMs classifier, which we proposed as part of our solution for the 2018 edition of Interspeech Computational Paralinguistics Challenge [24]. Here, our aim is to demonstrate the efficacy of GEWELMs for the task of predicting severity of mania.

Greedy Ensembles of Weighted Extreme Learning Machines

An Extreme Learning Machine (ELM) is essentially a single layer feed-forward neural network where the hidden layer is assigned randomly generated weights which are not updated during the training process. For classification, the output from the hidden layer can be mapped to the training labels using a least squares regression [122]. The idea is that even with random weights, the hidden layer can learn useful representation of input data which can be exploited by designing a suitable output layer. An outstanding advantage of ELMs is their very fast training time, which eases the process of tuning the hyper-parameters and experimentation.

We note that while ELMs were popularised by Huang et al. in 2004 [122], the fundamental concepts of ELMs have existed for much longer. Ping et al. proposed using least squares regression to compute weights of a neural network in [333]. The *random weights* concept of ELMs is analogous to the concept of random projections for feature mapping. If the number of neurons in the hidden layers is smaller than dimensionality of the input data, the ELM essentially implements dimensionality reduction. Conversely, when the number of neurons are larger than the input dimensions then the ELM performs dimensionality expansion.

The technique of dimensionality reduction using random projections is supported by the 1984 Johnson-Lindenstrauss Lemma [334], according to which *'points in a vector space of sufficiently high dimension, may be projected into a suitable lower-dimensional space in a way which approximately preserves the distances between the points'*. Meanwhile, dimensionality expansion is supported by Cover's theorem [335], according to which *'a complex pattern-classification problem, cast in a high-dimensional space non-linearly, is more likely to be linearly separable than in a low-dimensional space, provided that the space is not densely populated'*.

In our work, we use ELMs as a method for dimensionality reduction followed by least squares regression towards class label prediction. As such, we do not use a non-linear activation function. Moreover, we use principal component analysis (PCA) to decorrelate features prior to using ELMs.

It is important to note that ELMs have previously been reported to provide good performance for tasks pertaining to emotion recognition [317], Interspeech ComParE [112, 318], and depression recognition [128]. Furthermore, to deal with class imbalance in datasets (which also exists in AVEC 2018), Zong et al. [322] assign weights to each class according to the number of training examples available for that class.

A typical WELM classifier has two hyper-parameters: (1) the number of neurons L in the hidden layer and (2) the regularisation parameter c required for the generalised Moore-Penrose inverse [322], which can be tuned. In our work, we fix $C = 1$, and use four values for L i.e. $L \in \{2, 5, 10, 20\}$. WELMs have recently been proposed for tasks pertaining to social signal processing [37, 161].

It is quite obvious that not all random projections will yield acceptable results in terms of UAR for the classification tasks at hand. Some random projection vectors may actually reduce the separability between classes, while others may increase the separability. Rather than manually sift for useful random projection vectors, in this work, we propose *Greedy Ensembles of Weighted Extreme Learning Machines* (GEWELMs).

The fundamental idea behind GEWELMs is to train a sufficiently large number of WELMs and then select those which have UAR above a certain threshold for the development partition. We arbitrarily fix the threshold as the value corresponding to 90th percentile of the UAR of all WELMs in the ensemble. We do appreciate the fact that GEWELMs can have a tendency to over-fit to the development partition, hence we train two sets of GEWELMs. The first regime is called T2D-GEWELMs, which is the conventional training on the training partition and testing on the development partition, and the second is called D2T-GEWELMs, where we train on the development partition and test on the training partition. This serves to regularise the selection of WELMs in the ensemble by mandating that the set of random projections used for a particular WELM have acceptable performance for both T2D-GEWELMs and D2T-GEWELMs.

5.7 Experimental Results and Discussion

We perform various experiments in order to ascertain the efficacy of methods proposed in previous section. These experiments are detailed in the following sub-sections.

5.7.1 Session-wise Classification

The first experiment we perform is to investigate whether it is better to perform classification over the entire recording as a single entity or to classify each session independently and later perform fusion to yield a label for the recording.

To this end, we set up the following experiment. We compute audio features using the four feature sets from the openSmile toolkit [30], as discussed in

Table 5.2: UAR (%) achieved for classification on the development partition for four standard audio features using LIBLINEAR, with features computed (a) session-wise and (b) entire recording as a single entity

Feature Set	**UAR (%)**	
	entire recording	session-wise
`Prosody`	33.33	48.41
`IS10-Paraling.`	38.89	48.15
`ComParE`	39.95	48.15
`eGeMAPS`	36.51	49.74

Section 5.6.3, and perform classification. These features are computed, both, over the entire recording as a whole, and for each session of the recording. Finally, classification is performed using grid search for solver and cost using the LIBLINEAR toolkit. As clearly evident from Table 5.2, the classification accuracy is higher when classification is performed for each session and fusion is performed to yield labels for the recording. In fact, when prosody features are computed whilst considering the entire recording as a single entity, the UAR is at chance level. This result makes sense given recordings are well over a minute, while behaviour typically changes over a much smaller time duration [34, 152].

5.7.2 Turbulence Features for Audio/Visual Modalities

The second experiment we perform is on the efficacy of turbulence features discussed in Section 5.6.1 to capture the differences in feature trajectories of both audio and visual features.

From visual modality, we seek to capture both movement and emotional changes. For this we use the horizontal and vertical eye-gaze angle, the three-dimensional head-pose, and strength of six Facial Action Units (FAUs) computed via the OpenFace toolkit v2.0 [32]. These FAUs include: (1) `AU02_r`, which is the outer brow raiser, (2) `AU05_r`, which is the upper lid raiser, (3) `AU06_r`, which is the cheek raiser, (4) `AU12_r`, which is the lip corner puller, (5) `AU20_r`, which is the lip stretcher, and (6) `AU45_r`, which represents blinking. Early experimentation suggested that `AU06_r` and `AU12_r` were not useful for the task at hand, therefore, we did not consider them for subsequent experimentation. We must mention here that our aim is not to rule out the usefulness of these FAUs altogether, and may perform further experiments at a later date.

From audio modality, we compute turbulence features over contours of pitch (F0), and the first five formants (F1–F5). Our inspiration for using pitch was based on the success of turbulence features for pitch contours from the

Table 5.3: UAR (%) achieved for classification on the development partition for turbulence features computed for feature contours of visual features

Features	GEWELMs		LIBLINEAR
	per session	aggregate	aggregate
AU_02r	43.06	48.70	48.68
AU_05r	42.18	44.18	44.97
AU_20r	42.35	53.97	48.41
AU_45r	41.06	42.59	37.30
gaze angle (x)	46.25	46.03	45.77
gaze angle (y)	39.95	42.33	47.62
Pose (Rx)	41.60	47.62	44.44
Pose (Ry)	46.64	47.35	50.26
Pose (Rz)	45.18	48.94	47.35

Table 5.4: UAR (%) achieved for classification on the development partition for turbulence features computed for feature contours of audio features

Features	GEWELMs		LIBLINEAR
	per session	aggregate	aggregate
F0	40.68	48.94	45.77
F1	41.52	53.44	45.22
F2	38.42	49.74	36.77
F3	41.36	52.38	39.15
F4	39.03	42.33	46.03
F5	42.84	49.47	44.71

depression sub-challenge from AVEC 2017 [34], whereas formants were deemed useful for classification for bipolar disorder in [184].

Performance of turbulence features computed for audio/visual modalities are summarised in Table 5.3 (visual modality) and Table 5.4 (audio modality). In addition to the LIBLINEAR toolkit, we also perform classification using GEWELMS (see Section 5.6.4 for details).

5.7.3 Fisher Vector encoding of ComParE LLDs

The next experiment we performed is based on Fisher Vector encoding of ComParE LLDs (see Section 5.6.2 for details). In order to investigate the impact of the number of GMM clusters, we compute FVs for six values of GMM clusters i.e. 16, 32, 48, 96, 128, 192. Classification accuracy achieved with FV features is summarised in Table 5.5, where one can observe that this approach yields the highest UAR amongst all other features. These results highlight

Table 5.5: UAR (%) achieved for classification on the development partition for Fisher Vector features

# Clusters	GEWELMs		LIBLINEAR
	per session	aggregate	aggregate
16	46.07	53.97	55.56
32	48.08	62.17	56.08
48	48.23	59.26	55.56
96	45.74	55.82	57.41
128	45.27	51.85	59.79
192	48.47	63.49	55.03

Table 5.6: Comparison of UAR(%) achieved for classification on the development partition using GEWELMs and LIBLINEAR toolkit for standard feature sets

Features	GEWELMs		LIBLINEAR
	per session	aggregate	aggregate
`Prosody`	47.06	50.00	48.41
`IS10-Paraling.`	42.23	55.29	48.15
`ComParE`	43.35	54.23	48.15
`eGeMAPS`	48.65	51.85	49.74

two important points: (1) ComParE LLDs contain valuable information which is relevant for screening of bipolar disorder from acoustics of speech, and (2) Fisher Vector encoding of ComParE LLDs is a viable approach for classification of bipolar disorder from speech (at least for the given dataset). Furthermore, while we do not observe a correlation between number of GMM clusters and classification accuracy for GEWELMs, there is generally steady increase in performance for classifiers trained using the LIBLINEAR toolkit.

5.7.4 Classification using openSmile feature sets

Our final experiment on the development partition is based on computing classification accuracy using each of the four standard-feature sets provided with the openSmile toolkit. This experiment is an extension to results discussed in Table 5.2. Here our aim is to compare the classification performance achieved via GEWELMs with that achieved using the LIBLINEAR toolkit. One can clearly note that GEWELMs consistently perform better for all cases, in particular `IS10-Paraling.` and `eGeMAPS` feature sets.

Table 5.7: Summary of baseline and proposed methods on the test partition

Features	Modality	UAR
Baseline Methods		
`eGeMAPS`	A	50.00
`DeepSpectrum`	A	44.44
`FAUs`	V	46.30
`eGeMAPS` + FAUs	A+V	57.41
`DeepSpectrum + FAUs`	A+V	44.44
Proposed Methods		
`attempt 1`	V	57.41
`attempt 2`	A	42.59
`attempt 3`	A	48.15
`attempt 4`	A+V	51.85
`attempt 5`	A+V	46.30

5.7.5 Results on the Test Partition

Organisers of the AVEC 2018 workshop kept the labels for the test partition to themselves so that participants can develop methods for automated screening of bipolar disorder using the training and development partitions, and test the efficacy of their proposed methods on the test partition — without the bias of overfitting the test partition. Each participating team was allocated a total of five attempts at predicting the labels of the test partition.

In Table 5.7, we have summarised the results of our proposed methods for screening of bipolar disorder using both, audio and visual modalities. Our first attempt was majority-voting based fusion of turbulence features computed for feature trajectories of facial features. We used all features mentioned in Table 5.3 except `AU_45r` (which represents blinking), since the UAR achieved through this feature on the development partition was quite small. On the test partition, we achieved a UAR = 57.41% whilst using these features. This matches the baseline UAR for the challenge, which was achieved after fusion of audio-visual features which achieved best performance on the development partition. It is important to mention here that for visual modality only, our proposed features beat the baseline by a difference of UAR = 11.11% (i.e. 57.41% to 46.30%), which amounts to an improvement of more than 19%.

Our second attempt, however, was not as fruitful when it comes to accuracy on the test partition. Here, we submitted predictions on the test partition based on turbulence features for pitch and formant frequencies. While these features have a reasonable performance on the development partition, as shown in Table 5.4, these features achieved a UAR = 42.59% on the test partition.

Our third attempt on the test partition was based on predictions achieved

using Fisher vector features. On the development partition alone, these features achieved best results for both GEWELMs and LIBLINEAR toolkit. However, rather than simply choosing a single best performing model, we chose to use majority vote based fusion of predictions obtained through both GEWELMs and LIBLINEAR toolkit with the condition that performance on the development partition be greater than or equal to UAR = 55.00% (this threshold was chosen arbitrarily). Using this approach, we achieved a UAR = 48.15% on the test partition. While this result is better than our second attempt, we do note a significant drop in the performance of these features from the development to the test partition, which is likely due to overfitting on the development partition. Nevertheless, the performance is still better than the baseline DeepSpectrum features [336], which achieved a UAR = 58.20% on the development partition, which decreased to UAR = 44.44% on the test partition. It could be that there exists a difference between characteristics of patients with bipolar disorder within the development and test partitions.

Our fourth attempt on the test partition included majority-vote based fusion of predictions obtained from the following features: (1) FV 32-GMM with LIBLINEAR toolkit classifiers, (2) FV 32-GMM with GEWELMs, (3) predictions from the first submission which achieved UAR = 57.41%, (4) eGeMAPS with LIBLINEAR (baseline), and (5) FAU features with LIBLINEAR toolkit classifier (baseline). Our motivation to use the features from the baseline paper was that these features have been reported to produce highest performance, and when combined with our features could increase the UAR. On the test partition, we achieved a UAR = 51.85% which is better than that achieved with 2nd and 3rd attempts, but is still smaller than our best result on the test partition i.e. UAR = 57.41%.

For our final attempt on the test partition, we perform majority fusion on predictions achieved via FV 32-GMM with GEWELMs, predictions from the first submission, and the result of probability-based fusion of predictions eGeMAPS and FAUs from the baseline paper. Thus we combine our best result on the test partition, the best result for the test partition as per the baseline paper, and our best result on the development partition. However, we only achieved a UAR = 46.30% for the test partition, which is far below our expectations given that predictions were majority fusion of best results on the development and test partitions.

Given limited attempts on the test partition, we could not identify the cause of overfitting on the development partition, although we posit that there are likely to be some confounding factors which influence our machine learning models.

5.7.6 Comparison with submissions from other researchers

As per Ringeval et al. [337], in total 41 teams participated in the AVEC 2018 Bipolar Disorder challenge, 11 teams submitting results, and 4 papers were

accepted for oral presentation at the AVEC workshop. Our paper was amongst the 4 papers to be accepted. Other researchers whose papers were accepted include Du et al. [338], Xing et al. [339], and Yang et al. [340]. In subsequent paragraphs, we provide a summary of the work carried out by these researchers and end this section by comparing the final results of the AVEC 2018 Bipolar Disorder challenge.

Du et al. [338] propose a deep learning based system called IncepLSTM to capture multi-scale temporal information from acoustic LLDs for the task of predicting severity of bipolar disorder. IncepLSTM essentially integrates a CNN based Inception module [89, 341, 342] and LSTM [343]. We find it interesting that Du et al. mention our previous work on depression recognition [34] (and Section 4.7.1 as their inspiration for proposing a deep learning based solution for multi-scale aggregation of temporal information. In addition to the proposed IncepLSTM, they also propose a severity-sensitivity loss, inspired from the triplet loss [91], for the task at hand. The severity-sensitivity loss aims to jointly optimise the cost of minimising the distance between examples from the same class and maximising the distance between different classes. They compare the performance of their proposed system against three baselines: (a) SVM classifier learnt using MFCCs LLDs along with its velocity and acceleration contours, (b) SVM classifier learnt using eGeMAPS features [58], and (c) SVM classifier learnt using DeepSpectum features [336]. As per results reported in their paper, IncepLSTM achieved better UAR on the development partition as compared to the three baselines. These results suggest that CNN based inception module cascaded with LSTM is a powerful model to capture the temporal information from acoustic LLDs. However, it is also important to mention here that Du et al. did not report results on the test partition which would have provided an unbiased evidence about the efficacy of their proposed system.

Xing et al. [339] base their work on the previous work of Gong et al. [131] who had demonstrated that audio, visual, and text features sorted according to the context of spoken language are useful for the task of automated screening of depression. However, since organisers of AVEC 2018 challenges did not provide interview transcripts as part of the dataset, Xing et al. used Google Cloud Platform (GCP) to generate transcripts for interview sessions. Next, using these transcripts, they organised audio/visual recordings into three sets i.e positive, negative, and neutral, on the basis of the valence of spoken language. Similar to Gong et al., Xing et al. then computed a large set of audio, visual, and text features. For the audio modality, they computed eGeMAPS and MFCC features, for the visual modality, they computed MHH based histograms [300] for action units as well as evidence of Ekman's seven basic emotions which they computed using the Faceplusplus toolkit [344]. Finally, for the text modality, they computed a set of linguistic features from the SALAT toolkit [309], for which they sought inspiration from previous work by Dang et al. [158] who used linguistic features from the SALAT toolkit for the task of automated

depression screening.

Given that Xing et al. computed a large number of features, they used Analysis of Variance (ANOVA) [35] for feature selection in order to identify most useful features for the task at hand. Finally, they used eXtreme Gradient Boosting (XGBoost) [345], an algorithm which has been successfully used for various data science competitions, for the purpose of classification. However, rather than using XGBoost directly for the tertiary classification task, they use a hierarchical approach where the tertiary classification task is simplified into pairs of binary classification tasks. As per results reported in their paper, they achieve an impressive UAR = 86.77% on the development partition. However, the accuracy on the test partition drops down to 57.41%. The difference between the accuracies on the development and test partitions suggest that their models overfit the development partition. This is not uncommon when feature selection is used to optimise performance directly for the development partition without using a separate hold out partition.

The third paper accepted for the AVEC 2018 Bipolar Disorder challenge was the work of Yang et al. [340]. They proposed histogram based arousal features for 'bipolar depression classification'. It is important to mention here that depression was amongst the exclusion criteria when subjects were recruited for AVEC 2018 Bipolar Disorder dataset. Thus it was ensured that no subject in the dataset suffers from depression.

Nevertheless, the idea behind their proposed approach is quite interesting. They hypothesise that individuals with different severity of bipolar disorder have varying degrees of arousal and surmise that features derived from their arousal score can be used to screen for bipolar disorder. However, since arousal scores were not provided as part of the dataset, they had to first train a model to predict arousal of subjects in the AVEC 2018 dataset. To this end, they trained the LSTM-RNN model of He et al. [346] on the AVEC 2015 Affective dataset [347] and then used the trained model to predict arousal scores of subjects in the AVEC 2018 Bipolar Disorder dataset. These arousal scores were then aggregated using histograms to yield a fixed length global representation of the arousal scores for the entire recording.

Similar to their previous work [306], they also computed a large number of features from audio/visual modalities including audio features using the OpenSmile toolkit [30], visual features to capture body movement using the OpenPose [348], and facial action units. Given the large number of features, they used correlation based feature selection [349] along with brute force sequential forward search algorithm [350] and an SVM classifier to reduce the dimensionality of their feature set. Finally, they used DNN and Random Forest classifier based model fusion framework for classifying subjects according to the severity of their mania. As per results reported in their paper, they achieved a UAR = 71.41% on the development partition and a UAR = 57.41% on the test partition.

As it is obvious from the survey of publications from the AVEC 2018

Table 5.8: Summary of results for AVEC 2018 Bipolar Disorder Challenge

References	**UAR**	
	Dev	Test
Ringeval et al. [39] (*baseline*)	63.50	57.41
Du et al. [338]	65.10	–
Xing et al. [339]	86.77	57.41
Yang et al. [340]	71.41	57.41
Syed et al. [351]	–	57.41

Bipolar Disorder challenge provided in this section, the ceiling UAR on the test appears to be 57.41%. The baseline paper of Ringeval et al. [39] as well as work of three participants of the challenge (including us) were able to match the challenge baseline on the test partition but not beat it. This observation is certainly surprising since our work as well as the works of Xing et al. and Yang et al. is fundamentally different. Our result of 57.41% on the test partition was obtained via turbulence features computed for facial action units, Xing et al. compute a large number of audio, visual, and text features along with feature selection gradient boosting based classification to achieve the same performance test, and Yang et al. used a large set of features along with feature selection and an ensemble of DNN and Random Forest classifier to achieve the same score.

We surmise that the ceiling value for test partition UAR could be because some subjects in the test partition have grossly different behavioural characteristics which cannot be learned from the subjects in the training and development partitions. Nevertheless, as it stands, our work matches the state of art for the AVEC 2018 Bipolar Disorder challenges.

5.8 Time-Complexity Analysis

In this section, we provide an analysis of computational complexity of our proposed methods in terms of *time-complexity*. Time-complexity is defined as the computational complexity which describes the amount of time it takes to execute an algorithm [352].

In order to quantify time-complexity for our proposed algorithms, we measured their run-time using Matlab's *tic*/*toc* functions which provide the elapsed time during code execution. All time-complexity experiments reported in this section were performed on a personal computer with the following specifications: Intel Core i7-4790 CPU @ 3.60 GHz, 32.0 GB RAM, Windows 7 OS, and running Matlab 2018a. While these experiments were being conducted, no other simultaneous task was being performed on the PC, except standard background processes of Windows 7. Therefore, it is important to emphasise

that while this method can provide meaningful analysis into time-complexity, it is a crude method nonetheless.

We report time-complexity of four algorithms used in this thesis, namely, FV encoding of acoustic LLDs, turbulence features, LibLinear classifier, and GEWELMs classifier. It is pertinent to mention here that for FV encoding we used the Vlfeat toolkit [321], whereas for LibLinear, we used the LibLinear toolkit [171]. Both of these toolkits offer highly optimise Matlab executable (mex) files for the task at hand. Meanwhile, for turbulence features and GEWELMS classifier we wrote our code in Matlab.

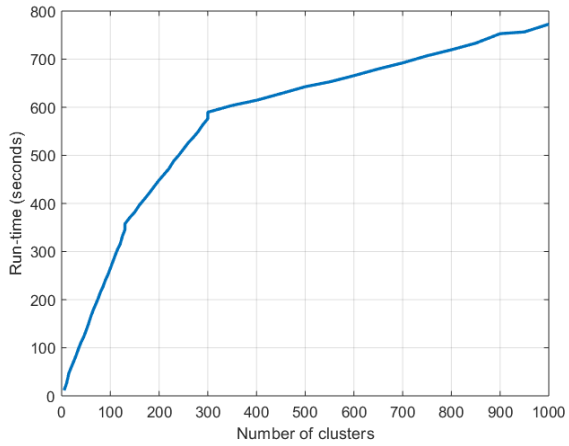
We trained GMM with acoustic LLDs from the ComParE feature set, with the number of clusters increased iteratively from 5 to 1000. Each GMM was allowed to train for up to a maximum of 100 iterations. For computing time-complexity of turbulence features, we took pitch feature as an example. Meanwhile, we used the same parameters for LibLinear and GEWELMS are reported in Section 5.6.4.

We start with the time-complexity for FV features. As previous discussed in Section 4.7.3, the generation of FV features consists of two main parts. The first part constitutes training a generative background model based on the GMM, whereas the second part involves computation of FV features. Therefore, we shall report the time-complexity of each of these parts separately.

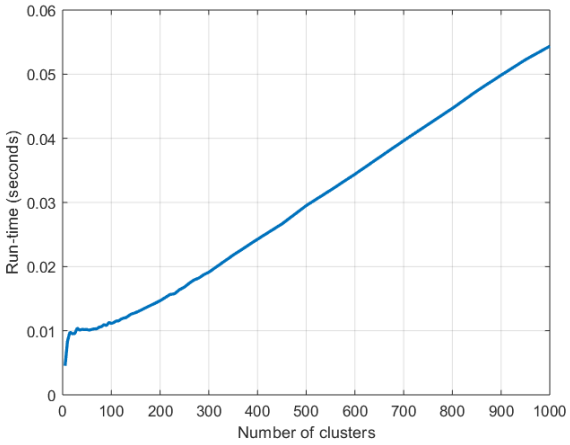
The time-complexity for the process of training GMMs is illustrated in Figure 5.1a, where we plot the run-times against the number of clusters for the GMM. Here, one can note that the time-complexity curve follows a logarithmic trend which means that order of time-complexity is also logarithmic with respect to the number of clusters. The time-complexity for the process of computing FV features is illustrated in Figure 5.1b, where, again, we compare the run-times against the number of clusters. Here, one can note that the time-complexity curve follows a linear trend which means that order of time-complexity is linear. It is important to mention here that the run-times shown in Figure 5.1b is the average run-time of 22444 samples.

In Figure 5.2, we plot the time-complexity for LibLinear and GEWELMs classifiers for FV features as function of the number of clusters for GMM. While the run-time curves for LibLinear and GEWELMs follow a logarithmic trend with respect to the number of clusters, one can note that GEWELMs requires considerably less run-time as compared to LibLinear. In order to compare the run-times for these classifiers, we also plotted their them against each other and found that GEWELMs is approximately 15 times faster than LibLinear.

Finally, in Figure 5.3, we plot the time-complexity for turbulence features. In addition to the AVEC 2018 challenge on bipolar disorder recognition, the multi-resolution feature aggregation approach offered by turbulence features was also used for AVEC 2017 challenge on depression severity prediction. Given that turbulence features are computed for feature contours, we provide time-complexity as a function of the length of feature contour. The curve

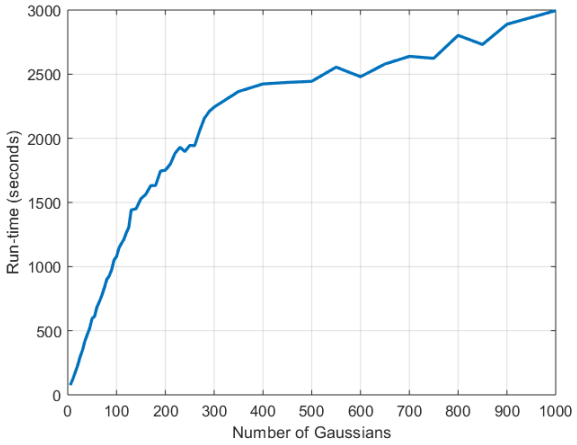


(a) Run-time-complexity for training GMM models

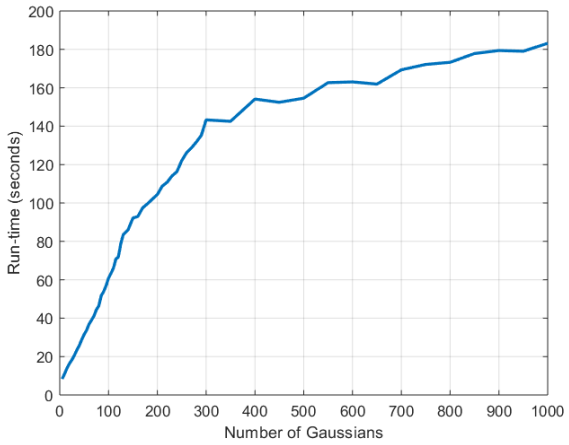


(b) Run-time-complexity for the process of computing FV features

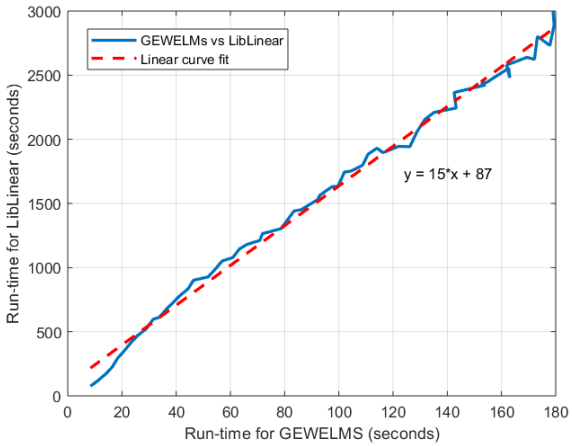
Figure 5.1: Run-time-complexity for training GMMs and computing FV features as a function of the number of clusters for the GMM



(a) Time-complexity for LibLinear



(b) Time-complexity for GEWELMs



(c) Ratio of time-complexities for LibLinear and GEWELMs

Figure 5.2: Run-time-complexity for LibLinear and GEWELMs classifiers

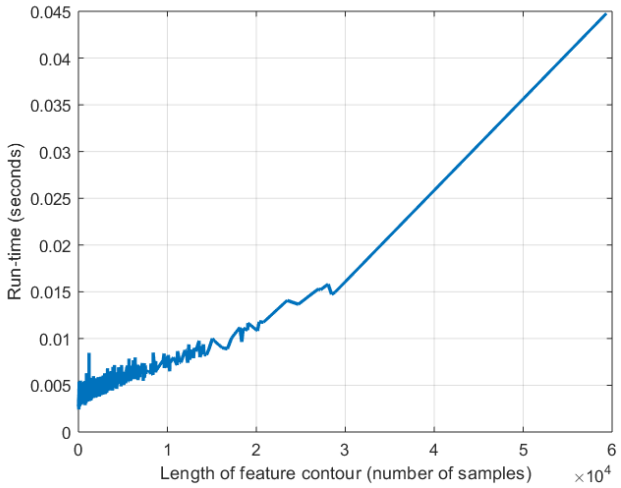


Figure 5.3: Run-time-complexity for turbulence features as a function of the length of feature contour

in Figure 5.3 shows that computation of turbulence feature requires small run-times (in the order of milliseconds), which highlights the advantage of these features. Furthermore, we note that the time-complexity of turbulence features has approximately linear characteristics. Similar to the process used for FV features, we also computed the run-times for LibLinear and GEWELMs classifier when turbulence features were provided as input. Unlike FV features, where the dimensionality of feature vectors grew linearly with the number of GMM clusters, the dimensionality of the feature vector for turbulence features is always fixed to 45. We report that the run-time for LibLinear classifier for turbulence features is 6.20 seconds whereas the run-time for GEWELMs classifier is 4.44 seconds.

5.9 Towards Automated Clustering for FV features

So far in our work, we have used cross-validation accuracy on the development partition in order to determine the optimal number of clusters. However, as clear from Figure 5.2, cross-validation can be time-consuming process. Therefore, in this section, we seek to investigate the feasibility of selecting the optimal number of clusters for FV features based on how well the GMM represents the training data. While doing so, we would still like to pursue our main objective i.e. to achieve maximum cross-validation accuracy on the development partition (so that it can also generalise on to the test partition).

To this end, we shall use the log-likelihood value of the GMM as a measure

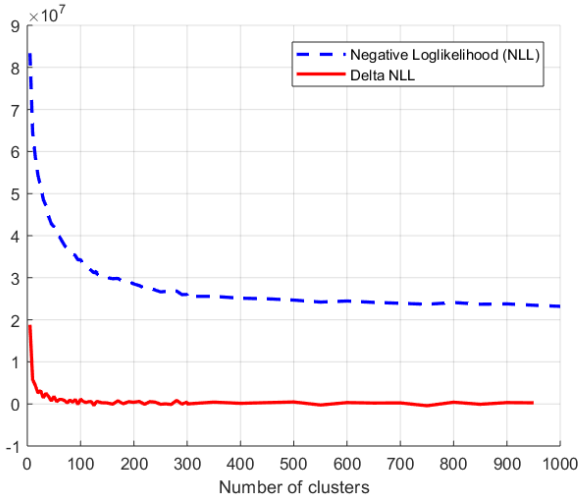


Figure 5.4: Negative-loglikelihood (NLL) and delta-NLL for GMM models with various cluster sizes

of about how well the GMM fits the training data [95]. The objective while training the GMM is to optimise the mean, covariance, and mixture weights for the GMM such that the log-likelihood (LL) value is maximised or conversely, the negative log-likelihood (NLL) value is minimised.

Similar to the experiments conducted for time-complexity analyses, we train GMM with acoustic LLDs from the ComParE feature set, with the number of clusters increased iteratively from 5 to 1000. Each GMM was allowed to train for up to a maximum of 100 iterations. Training of GMMs and the computation of FV features was performed using the Vlfeat toolkit, and both LibLinear and GEWELMs classifiers were used to compute the cross-validation accuracy in terms of UAR on the development partition.

In Figure 5.4, we plot the NLL for GMM against the number of clusters. It is clear that the curve contains three different segments. The first segment, to the far left, contains a steep downward slope for NLL which represents the improvement in the GMM's ability to represent its training data. The second segment, located between the regions of steep downward slope and the flat region is a transitionary in which improvement in the performance of GMM provides diminishing returns. Note that increase the number of clusters comes at the cost of increasing time-complexity, as discussed in previous section. The third segment of the curve consists of the region where there is little to no improvement in the NLL.

In order to gauge the improvement with respect to NLL as a function of the number of clusters, we compute change in NLL between adjacent points on the x-axis (i.e. the cluster number) based on the work of [353], and plot this in Figure 5.4. Here, it is clear that increasing the number of clusters beyond

80 offers only negligible improvement in the NLL. Moving forward, we shall now limit our analysis to a maximum of 200 clusters rather than a maximum of 1000 clusters and conduct further investigation.

In addition to NLL, it is also possible to select the number of clusters based on the methods of Akaike information criterion (AIC) [354]. AIC penalises the NLL of models based on their computational complexity by a factor equivalent to $2k$ to the NLL, where k is the number of parameters in the GMM. In Figure 5.5, we show results for NLL and AIC along with their deltas for GMMs with number of clusters between 5 and 200. It is apparent that NLL and AIC have similar interpretations.

Finally, in Figure 5.6, we plot the cross-validation UAR on the development partition for FV features using both LibLinear and GEWELMs classifiers. Here one can note that the UAR generally improves for up to 80 clusters, which is in line with our understanding from cluster selection based on NLL. There are, however, considerable variations in the cross-validation UAR which are expected – we are essentially optimising the number of clusters for GMM as well as the parameters of LibLinear and GEWELMs classifiers on previously unseen data. Nevertheless, our method for automatic clustering can be used to reduce some of time-complexity for features based on Fisher Vector encoding.

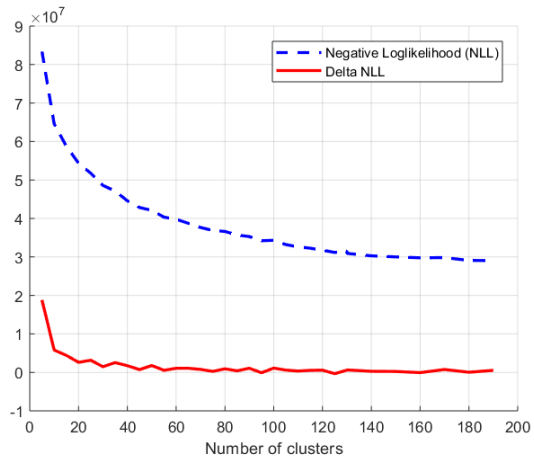
5.10 Summary

In this chapter we introduced two new approaches for the task of automated screening of bipolar disorder from audio-visual recordings; namely turbulence features and Fisher Vector encoding of ComParE LLDs. We also introduced GEWELMs based classification of these features and demonstrated the efficacy of these methods on both, the development and test partitions.

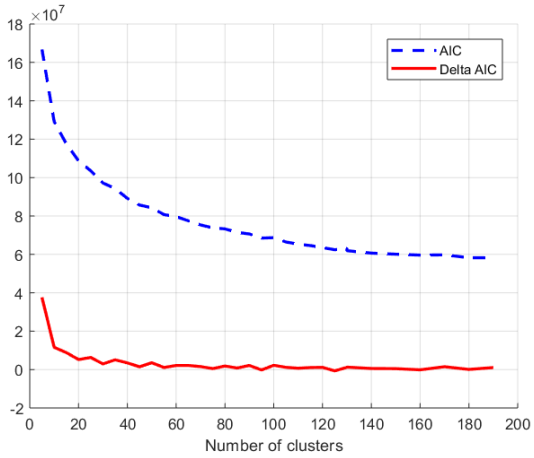
To conclude, we summarise the key contributions of our work as follows:

- We report that for the task of predicting severity of mania, *turbulence features* computed for visual modality performed better than the audio modality. In fact, the best result achieved by us on the test partition i.e. UAR = 57.41% uses turbulence features for visual modality. This result exactly matches the best result published as the official baseline, which was a result of fusion of features from audio and visual modalities.

- Fisher Vector encoding of ComParE LLDs achieved best performance in terms of classification accuracy amongst all other features on the development partition. However, their superior performance could not be replicated on the test partition, likely because of overfitting on the development partition. Given limited attempts on the test partition, we could not identify the cause of overfitting on the development partition, although we posit that there are likely to be some confounding factors which influence our machine learning models.



(a) NLL and delta-NLL



(b) AIC and delta-AIC

Figure 5.5: Model fitting measures for GMM models with various cluster sizes

- We investigated the efficacy of four standard feature sets from the openSmile toolkit i.e. Prosody, IS10-Paralinguistics, ComParE functionals, and eGeMAPS features. We found IS10-Paralinguististics and eGeMAPS feature sets to be most useful. It is important to mention here that while organisers report that eGeMAPS features achieve a UAR = 55.03% on the development partition [39], we could not replicate this result, even though we used same experimental settings as reported by them.

- We also investigated whether it is better in terms of accuracy to perform classification over the entire recording as a single entity or to classify each session independently and later perform fusion to yield a label for the recording. Based on our experiments, we report that session

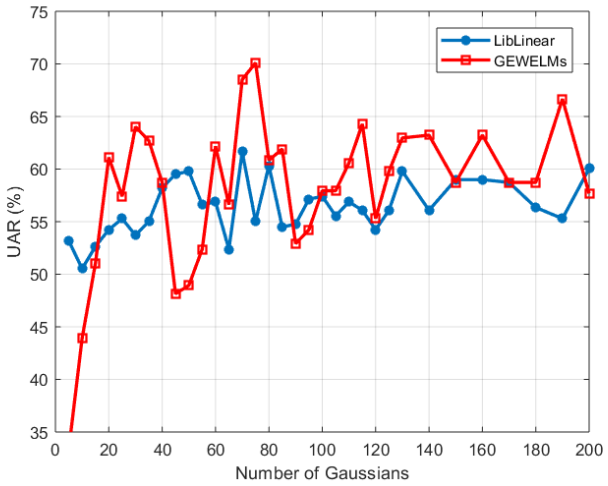


Figure 5.6: Cross-validation UAR with LibLinear and GEWELMs classifiers

based classification is a better option when it comes to classification accuracy. This is somewhat contrary to the findings of Ciftci et al. [33], who reported no advantage of similar segmentation.

- We report the time-complexity of our proposed methods i.e. FV features for acoustic LLDs, turbulence features for acoustic and visual LLDs, and GEWELMs.

- We proposed a semi-automated method for shortlisting a range of potentially useful clusters based on the negative log-likelihood (NLL) and Akaike information criterion (AIC) values for the GMM. This method provides an alternate to cluster selection based on the cross-validation accuracy on the development partition.

- In our attempt of crafting features based on background knowledge of bipolar disorder, we found that certain aspects of behaviour of subjects cannot be probed directly from audio-visual recordings. For example, lack of requirement for sleep is a key behavioural indicator for individuals with mania as per the YMRS [22,23]. Now, unless the subject is explicitly asked a question about their sleeping habits, it may not be possible to ascertain how much sleep a particular subject has been having. We attempted to quantify sleepiness using AU45, which represents blinking, but this approach performed poorly.

  Similarly, sexual activity/interest is another aspect of the YMRS which cannot be directly gauged from audio-visual recordings. For the AVEC 2018 BDS, we found that such questions were not asked by the subjects in the audio/visual recordings which are provided as part of the dataset.

While our aim was to propose features inspired from behavioural characteristics of individuals with mania as per the YMRS, it is necessary to acknowledge the existence of inherent limitations of this approach.

To conclude, we had proposed a number of methods for automated screening of bipolar disorder, in particular predicting severity of mania. We have successfully demonstrated the efficacy of these methods for the task at hand. While these results are competitive with respect to the AVEC 2018 Bipolar Disorder sub-challenge, we believe there is still a lot of room for improvement when it comes to classification accuracy.

Chapter 6

Automated Screening for Autism Spectrum Disorder

6.1 Introduction

The World Health Organisation (WHO) describes Autism Spectrum Disorder (ASD) as a range of behavioural conditions which lead to impaired social interaction and communication (both, verbal and non-verbal). Individuals with ASD typically have a narrow range of interests, and pursue activities which are odd, stereotyped, and generally repetitive [355]. These traits hinder social engagement, making life very difficult for people who have this disorder.

The Centers for Disease Control and Prevention (CDC) [356] report that about 1 in 59 children in the United States of America has been identified with ASD; boys are 4 times more likely to be affected by this disorder compared to girls. In the UK, these rates are much smaller, but still significant. According to the National Autistic Society, a charity for autistic people in the United Kingdom [357], there are around 700,000 people on the autism spectrum in the UK — that is more than 1 in 100 as per the 2011 UK census figures. ASD is reported to occur in all racial, ethnic, and socio-economic groups.

Although autism is incurable, early detection of ASD is critical for significant improvement in the quality of life for the affected individuals. According to the WHO [355], therapy sessions initiated while these individuals are still young can exploit the brain's plasticity to increase the chances of success at alleviating certain social engagement deficits. This is where automated screening can help, due to its potential advantages over conventional screening methods (as discussed previously in Section 1.4).

The rest of this chapter is organised as follows: we start with statements of novelty and contributions of our work on the development of automated methods for screening of ASD. We follow this with a literature survey to identify speech features which have high discriminative power when it comes to identifying individuals with ASD. We then describe the dataset which we

use in our work. This is followed by a discussion on a feature engineering and classification mechanism to screen for ASD using speech features. We then provide a discussion on the efficacy of standard feature sets for the task at hand. Finally, we end the chapter with a summary of contributions of our work.

6.2 Novelty and Contributions

Our work on the development of automated screening methods for Autism Spectrum Disorder (ASD) is novel in the sense that we performed numerous experiments and provide discussion on the effects of ASD on speech, in light of research literature, for a previously unpublished dataset. We started from raw audio/visual recordings of children with ASD as they talked with an interviewer and performed manual annotation for speech belonging to children so that it can be processed further.

The contributions of our work are listed as follows:

- We propose a feature engineering and classification mechanism for automated screening of ASD from speech for subjects in our dataset. The efficacy of this mechanism is demonstrated in terms of classification accuracy and the identification of highly discriminative speech features using Mann-Whitney U-test based statistical analysis [43] and effect size for statistically significant features [44, 45].

  The reader is referred to Section 6.5 for details about this contribution.

- We investigate the influence of speech segmentation on classification accuracy. To this end we performed experiments using six different segmentation rules. Our results suggest that classification accuracy is dependent on the duration of voiced speech in each speech segment.

  The reader is referred to Section 6.5 for details about this contribution.

- We report on the basis of our experiments that traditional voice quality features such as shimmer, jitter, and HNR are not able to provide discrimination between speech of individuals from TD and ASD groups. In addition, we report that features from the COVAREP voice quality feature set are able to discriminative between the speech of individuals from the two groups.

  The reader is referred to Section 6.6.3 and Section 6.6.4 about details of this contribution.

- We report on the basis of our experiments that spectral characteristics of speech individuals from TD and ASD groups are indeed different. This is an important observation because it suggests that individuals with ASD

may also suffer from changes at vocal tract level, similar to individuals with mental disorders.

The reader is referred to Section 6.5.5 and Section 6.6.3 about details of this contribution.

6.3 Literature Survey

According to the latest edition of Diagnostic and Statistical Manual of Mental Disorders (DSM-5) [18], ASD is a neuro-developmental disorder which is characterised by two key traits. The first trait of individuals with ASD includes deficiency in social interaction and communication skills, whereas the second trait is that these individuals also indulge in restricted and repetitive behaviours. As per the discussion provided in the work of Brukner-Wertman et al. [358], both of these traits need to be present before an individual can be diagnosed as having ASD.

A number of studies have investigated the differences between social signals of individuals with ASD and typically developing (TD) individuals in order to lay ground work for development of automated screening methods. These signals include facial expressions [68, 359], eye movement [360], body movement [361–364], and speech [141, 263, 276, 365–368].

In this thesis we explicitly focus on the speech of individuals with ASD. Our aim is to identify speech features, in light of published research literature, which have high discriminative power when it comes to identifying individuals with ASD from those who do not. Therefore, in line with our discussion on automated screening for disorders from audio modality (see Section 3.4), we undertake a literature survey in terms of three aspects of speech production i.e. prosody, voice quality, and spectral characteristics.

6.3.1 Prosody Analysis

Abnormal and disordered prosody is an established marker for communication deficits. Individuals with ASD are known to have speaking styles ranging from flat, monotonous, variable, 'sing-songy', and 'machine-like' to 'just bizarre' [46, 182, 369].

Fusaroli et al. [46], undertook a thorough and systematic survey of computational methods for screening of ASD, focussing on audio modality. In their survey, they report that only 2 out of 16 studies found average pitch to be statistically significant between TD and ASD group. Pitch variability (measured as range) was found to be statistically significant in 12 out of 22 studies; 11 of those reported pitch variability being larger for individuals with ASD, while only 1 reported a smaller pitch variability for the ASD group. Interestingly, DePape et al. [182] reported wider pitch for high functioning

ASD and narrower pitch for medium functioning ASD; implying that pitch variability may depend on subcategories of ASD.

Intensity is an important aspect of prosody and represents stress towards a particular utterance. Fusorali et al. [46] found the average intensity to be statistically significant for only 1 out of 8 studies, and the intensity variability to be statistically significant for 1 out of 2 studies, whereas all other studies reported null findings for intensity based features. The two studies which did report statistically significant results, found ASD group to have smaller intensity. It must be mentioned here that intensity measures are highly dependent on the relative position of the subject and the microphone, so intensity features may not be useful in practical cases where microphone placement is not consistent for all subjects in the study.

We also note from Table 1 in Fusaroli et al. [46] that speaking tasks for these experiments varied between various types of scripted and spontaneous speech, each of which can modulate prosody differently. Therefore, one needs to appreciate the heterogeneity of experiments in addition to any change in speech prosody due to ASD.

6.3.2 Voice Quality Analysis

Voice quality was considered as an extension of prosody [370], and is known to vary with respect to underlying emotional state [233, 371, 372]. Given that the ASD causes impaired behaviour and social communication, there is good reason to investigate the efficacy of voice quality analysis for screening ASD.

The winners of the Interspeech ComParE 2013 Autism sub-challenge, Asgari et al. [141] used HNR, jitter, and shimmer to classify between speech of typically developing (TD) and atypically developing (ATD) children. They did, however, use a harmonic modelling approach to recreate speech signal with minimal noise interference, prior to computing voice quality features i.e. they did not compute voice quality features for raw speech recordings.

Bone et al. [195] explored the hypothesis that voice quality features can be used to quantify perceptual depictions of odd voice quality for children with ASD. They used jitter, shimmer, HNR, and CPP as voice quality features, and report that median value of jitter had positive correlation with severity of ASD, whereas median HNR had negative correlation. Meanwhile, median value of CPP was not found to be significantly correlated with ASD severity.

While voice quality analysis offers interesting insights into changes in voice quality due to ASD, Fusaroli et al. [231] highlight that the use of non-standard acoustic feature descriptors for measuring voice quality has so far meant that one cannot be conclusive about the effects of ASD on voice quality.

6.3.3 Spectral Modelling

Spectral modelling approaches have been successful in recognising speech deficits of individuals with depression, as discussed in Section 3.4.3 . A number of researchers have surmised that speech impairments due to ASD can also be identified on the basis of spectral modelling of speech [141, 144, 145, 263, 270, 366, 367, 373–375].

To the best of our knowledge, amongst the first in this regard is the work of Bonneh et al. [366], who posit that speech abnormalities in ASD are reflected in its spectral content. They computed long term spectral average (LTAS) to capture various vocal tract configurations of subjects as they read scripted sentences. Their work does indeed show that TD and ASD groups have distinct spectra, with speech of the ASD group having a relatively *flat* magnitude spectrum. However, one needs to tread carefully while interpreting these results since TD and ASD groups in their dataset were recorded at multiple places, making LTAS prone to picking up channel characteristics in addition to any speech impairment due to ASD.

A number of researchers followed Bonneh et al.'s work and undertook experiments on spectral modelling. For example, Kakihara et al. [367] identify line spectral frequencies and MFCCs to provide enough discrimination between the TD and ASD groups to provide a classification accuracy of 80.20% on their dataset. Motlagh et al. [263] identify MFCCs, spectral centroid, and spectral roll-off amongst other features to be useful for classifying between speech of TD and ASD groups, achieving classification accuracy of 96.17% for their dataset.

Spectral modelling approaches were also explored for ComParE 2013 Autism sub-challenge [64]. Kirchoff et al. [373] used feature selection on 6,373-dimensional ComParE 2013 features [30], achieving better results than the official baseline for the challenge. They report that most of the top features provided information about speech spectra. Rasanen et al. [374], also used feature selection (albeit a different method) on ComParE features, and also report that spectral features were amongst the most discriminative. Meanwhile, Martinez et al. [375] use shifted delta coefficients spectral features proposed in [376] to capture long-term information of speech, although these features alone did not beat the official baseline.

The winners of the sub-challenge Asgari et al. [141] estimated parameters of a harmonic model of voiced speech and use it to reconstruct the speech signal in order to limit the amount of noise. Then from the reconstructed (essentially noise-free) speech signal, they quantified voice quality using HNR, shimmer, and jitter to achieve a UAR of 93.58% — the highest for the ComParE 2013 Autism sub-challenge.

Arguably the most important publication from the ComParE 2013 Autism sub-challenge was the work of Bone et al. [270]. They observed through informal listening that participants of the TD group in the corpus had reverberation

noise, and verified this by plotting LTAS of speech recordings (see Figure 1 in their paper). Next, they trained a single cluster GMM on the LTAS features targeting the range of frequencies over which reverberation was evident from the LTAS plots. They achieved an accuracy 79.70% for classifying between TD and ATD groups — high accuracy but below baseline, and 51.40% for classifying between TD and the three categories of ATD group — better than baseline. As it turned out [29], Bone et al. [270] were able to classify between TD and ATD groups by explicitly focussing on differences due to recording environment. We find the work from Bone et al. [270] significant because it exposed the pitfalls of blindly applying machine learning algorithms for clinical applications i.e. without taking efforts to establish a link between clinical knowledge and computer science experimentations. We revisit the significance of their work in Section 6.5.5.

More recently, Pokorny et al. [145] used the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) feature set [58] for classification between vocalisations of infants from TD and ASD groups. They list top features in terms of effect size computed for Mann-Whitney tests [43]. Amongst these, they report that six of the top ten features are spectral features, underpinning the potential efficacy of spectral features.

Baird et al. [144], computed a number of standard feature sets from previous ComParE challenges including IS09-emotion from ComParE 2009 [377], IS10-paralinguistics from ComParE 2010 [329], and ComParE 2013 features [64], and used linear SVMs [116] for classification. It is important to mention here that each of these feature sets contain a variety of spectral features. Baird et al. also used a deep learning framework based on CNNs to process spectrograms of speech recordings. By comparing results from Table 3 and Table 4 of their paper, one finds that every standard feature set performs better than the deep learning solution.

While it is evident from the surveyed literature that individuals with ASD do indeed have perceptually distinct characteristics to their speech, research so far has been inconclusive on what is the best way to quantify this distinctiveness: be it from prosody, voice quality, spectral modelling or a combination of these. Amongst these, speech prosody appears to be the most relevant aspect of speech production which can be probed to screening individuals with ASD. This is primarily because prosody is strongly correlated with communication skills [138], and impaired communication skills is known to exist in individuals with ASD [18].

We find no evidence — beyond experimental results on private datasets — for speech production at voice source and vocal tract level to be adversely affected by ASD. There is no reason, therefore, to suggest that voice quality and spectral modelling features should be useful. The DSM-5 manual [18] itself also does not provide guidelines which would support muscle control/motor changes in the vocal production system due to ASD, unlike, for example, depression which causes psychomotor changes. The DSM-5 manual explicitly focusses

on impaired social communication and restricted and repetitive behaviour as markers for ASD.

There also exists a lingering controversy that the dataset used for Autism sub-challenge at ComParE 2013, i.e. the Child Pathological Speech Database (CPSD), was affected channel noise due to different recording environment for typically developing and atypically developing groups [29, 195]. Results reported from the sub-challenge therefore need to be interpreted cautiously as they may well be compromised.

6.4 Dataset

We use the dataset collected by Dr. Catherine Jones from the School of Psychology, Cardiff University. It consists of audio/visual recordings of 49 school going children, out of which 27 belong to the TD group, with 17 males and 10 females. The ASD group consists of 22 subjects, with 18 males and 4 females. The mean and the standard deviation of the age for the TD group is 10.30 years and 0.88, respectively. The mean and the standard deviation of the age for the ASD group is 10.13 and 1.91, respectively.

The recordings assume two scenarios, both of which consist of spontaneous speech from subjects as they talk to the camera (they were instructed to do so). In order to ensure that there was enough speech data from each subject, as in [378] an interviewer is present behind the camera and intervenes only if a subject stops talking before at least two minutes worth of recording takes place. In the first scenario, the subjects describe their bedroom, whereas in the second scenario, subjects describe a cartoon that they have watched as part of the experiment.

The TD/ASD dataset is part of video recordings of these scenarios which have been annotated by the author of this thesis using ELAN software [41, 42], such that only sections of the interview process which contain the speech from the subjects were retained to create the dataset — video modality is not a part of the dataset. The dataset also does not contain any speech from the interviewer or very long silences during the interview.

Our experiments in this thesis are limited to this TD/ASD dataset, although we do draw from results and conclusions from published literature. Limitation in terms of dataset is because of ethical restrictions due to which we could not get access to datasets from at least three different research groups, i.e. works of Motlagh et al. [263], Bone et al. [195], and Baird et al. [144]. Similar ethical restrictions do not permit us to share our dataset either.

6.5 Screening for ASD from Speech

In this section, we discuss in detail our feature engineering and classification mechanism for automated screening for ASD from speech of subjects in our

dataset. Our proposed feature engineering mechanism consists of a four step process which includes speech segmentation, feature extraction, feature description, and feature selection. Since our dataset is rather small in size, we use 8-fold subject-dependent cross-validation, and report accuracy in terms of mean accuracy over the 8 folds. In order to describe the process flow we take the first cross-validation fold as an example, unless stated otherwise.

6.5.1 Speech Segmentation

Speech segmentation is a pre-processing step which is used to divide a continuous speech recording in such a way that relevant information for the task at hand is retained, while extraneous information is removed as much as possible. For example, for a speaker recognition task, speech from subjects may be the relevant information but background noise is extraneous information. While this pre-processing step is not mandatory, it can significantly improve results.

In order to convince the reader of the significance of speech segmentation for automated screening of ASD, we provide a brief overview of the three types of speech as a preamble to speech segmentation. Speech can be classified as either voiced, unvoiced, or silence. Voiced speech is a result of vibrations from the vocal cords and is harmonic in nature. Unvoiced speech also carries verbal information, but unlike voiced speech, vocal cords do not vibrate. It therefore does not include harmonic components. For example, in English language, the sound produced for consonant 'z' is voiced speech whereas the sound produced for consonant 's' is unvoiced speech [379]. Silence, meanwhile, does not include speech — neither voiced, nor unvoiced. It is still important for speech communication since it is used to convey prosodic information, and maintain speech intelligibility (long connected segments of speech may be unintelligible).

Nevertheless, for most practical applications, samples within speech recordings are classified as either voiced or unvoiced using a *voice activity detector* (VAD) [62, 380–382]. As noted by Koutrouvelis et al. [382], segments of speech recordings identified as *unvoiced*, do not strictly mean unvoiced speech, as these can also be *silence*. This is an important point to consider since one would not want to include too much of *silence* in speech recordings, as these features would only represent the recording environment, not a speech impairment due to, say, ASD.

To the best of our knowledge, there is no consensus on what is the best approach for segmentation. For example, Scherer et al. [173] create new segments for the DIAC-WOZ dataset if *silence* greater than 300 ms is detected. It is reminded for the reader's knowledge that the DIAC–WOZ dataset was used for AVEC 2016 DCC [12] and AVEC 2017 DSC [13]. Williamson et al. [128], the winners of the AVEC 2014 DSC [11], retained *silence* up to 750 ms. Meanwhile, data provided by the organisers of Interspeech Computational Paralinguistics Challenge (ComParE 2018) [148] for the Atypical Affect sub-

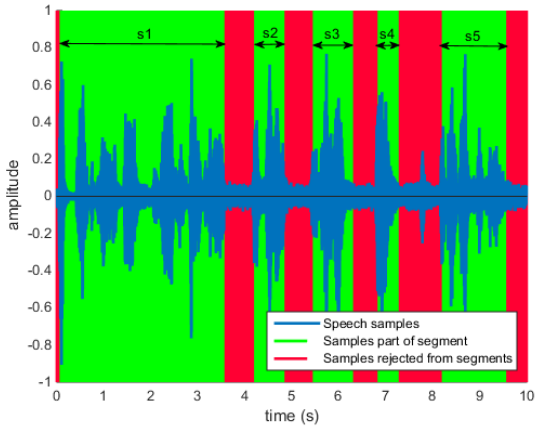


Figure 6.1: Illustration of speech segmentation, with $t = 150$ ms.

challenge performed 'manual' segmentation of speech recordings but did not elaborate further.

In this work, we perform experiments with six different segmentation rules, and report classification accuracy for each of these. Speech segmentation is implemented by pre-processing speech signals using the following rules depending on the duration of unvoiced speech/silence to retain.

> *Rule 1*: All contiguous voiced and unvoiced samples form a speech segment as long as the duration of unvoiced speech does not exceed t seconds (a tunable parameter), failing which a new speech segment is created.
>
> *Rule 2*: Every speech segment must have at least 100 ms of contiguous voiced samples.
>
> *Rule 3*: All unvoiced samples at the beginning and the end of a speech segment must be removed.

The distinction between voiced and unvoiced samples is made using decisions from a voice activity detector based on [383], provided as part of the VOICEBOX toolbox [384]. *Rule 1* controls the amount of unvoiced speech permitted within a speech segment. *Rule 2* ensures that a speech segment created through *Rule 1* comprises of at least four short time frames, each of which is 25 ms (typically used for short-time analysis of speech). Through experimental analysis we found that at least 100 ms of contiguous voiced samples are necessary to capture features which contribute to improvement in classification accuracy. Finally, *Rule 3* discards all unvoiced samples from either end of the speech segment, which may not have been removed by *Rules 1* and *2*.

An illustration of speech segmentation is given in Figure 6.1, where a sample speech signal acquired over a period of 10 seconds is divided into 5 segments. Each band in green represents a speech segment, created on the basis of the three rules. The bands in red represent samples which have been rejected. It is interesting to note that this particular setting for speech segmentation rejects some samples which may represent speech, but there aren't enough samples to satisfy *Rule 2*.

6.5.2 Feature Extraction

We compute various short-time features which describe temporal and spectral information of the speech signal (given below). It is important to mention here these experiments were completed prior to us adopting the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) feature set [58]. Our work was primarily motivated by the work of Alghowinem et al. [203, 224] on automated screening for depression.

Temporal features: ZCR, energy, energy entropy in time-domain

Spectral features: Chroma vector, formants, spectral centroid, spectral shape, spectral flux, HNR, energy entropy in frequency-domain, pitch, spectral roll-off, MFCCs and cepstral peak prominence (CPP)

Except for pitch and CPP, all features are computed over each short-time frame of 25 ms overlapped by a period of 10 ms. Pitch features are computed over a duration of 100 ms and CPP features every 10 ms which are in line with the default settings of the COVAREP toolbox [62]. All other features were computed using Audio Analysis Library [65], with default settings. In total, we computed 42 speech LLDs.

6.5.3 Feature Aggregation

The next step is to summarise information provided by LLDs mentioned in Section 6.5.2 in the form of derived features and aggregated features. This is achieved through a four stage feature description processes.

Stage 1 involves creating two sets of features. The first set is a copy of LLDs without any further processing on them. We name them `RAW` features. The second set consists of the z-scores values (mean centred and normalised by unit standard deviation) of the features for each segment. We call these Stage-1 features, and the process is repeated for all 42 features. For example, `CPP_RAW` is a Stage-1 feature.

At the second stage, we aim to extract dynamics of Stage-1 features which result due to changes in the characteristics of the speech signal. Therefore, we create three subsets for each of the raw and normalised sets from Stage-1. The first subset consists of features as they are i.e. without further processing.

The other two subsets contain the 1^{st} and 2^{nd} derivatives over time of the raw and normalised sets. Therefore we have 6 Stage-2 features for each of the 42 features i.e. raw, z-score of raw, 1^{st} derivative of raw, 2^{nd} derivative of raw, 1^{st} derivative of the z-score of raw and 2^{nd} derivative of the z-score of raw, and this means that we now have $42 \times 6 = 252$ Stage-2 features in total. An example of Stage-2 features is `CPP_RAW_2ndDeriv`.

In the third stage, we summarise the information described by Stage-2 features for each short-time frame within a segment into a single Stage-3 feature for that segment i.e. compute an aggregate feature. This is achieved by computing 13 statistical measures, which include mean, trimmed mean (at 10% trimming), standard deviation (StdDev), kurtosis, minimum, maximum, median, mode, range, inter-quartile range (IQR), dynamic range and the relative position of the peak value of the feature within the segment. Once completed, this process brings the total count for the number of Stage-3 features to 3276 i.e. $252 \times 13 = 3276$. An example of Stage-3 features is `CPP_RAW_2ndDeriv_StdDev`.

Finally, we use trimmed mean (at 10% trimming) to collapse each set of Stage-3 feature points for all speech segments into a single Stage-4 feature point.

6.5.4 Feature Selection

The machine learning task for training a classifier to distinguish between the speech of the TD and ASD groups is ill conditioned in our case. This is due to the relatively small size of the dataset i.e. 49 participants compared to the number of features i.e. 3276, which may lead to suboptimal training of the classifier and therefore result in poor classification results. We can, however, alleviate this problem to an extent by selecting a smaller number of features based on their ability to discriminate between the two groups, and then supply the selected features to the classifier.

The feature selection process we implement is based on the filter/wrapper approach discussed in [100]. Guyon et al. [100] suggest using a correlation metric to filter each feature on the basis of its discriminative power (univariate analysis). A wrapper based method is subsequently used to select the set of features which when used together can improve the classification accuracy beyond that of an individual feature (multivariate analysis).

As an example, we describe the process flow for feature selection for the first cross-validation fold as follows: for the filter method, we compute p-value from the pairwise t-test for each individual feature and use the p-value as the figure of merit for discrimination ability. While t-tests are typically used for statistical hypothesis testing, they can also be used for feature selection [224, 385]. Next, we sort features in ascending order with respect to their p-values, and retain features which have p-values less than 1.53×10^{-5}, while rejecting all other features. This threshold is computed by normalising the standard p-value cut-

Table 6.1: The effect of parameter t in Rule 1

Rule 1 parameter (t)	Classification accuracy	Mean dataset utilisation
∞	98.96%	100.00%
1000 ms	93.65%	90.87%
500 ms	88.35%	83.26%
200 ms	76.32%	77.22%
75 ms	83.14%	73.67%
25 ms	78.97%	72.58%

off i.e. 0.05 by the number of features i.e. 3276, thereby we apply Bonferroni Correction [386]. At the completion of the filtering step, we are left with 119 features from a total of 3276.

It is important to mention here that filter based methods only select features on the basis of their individual discrimination ability (univariate), and we are naturally interested in exploring the combinations of these features which can provide even better classification accuracy (multivariate). This is where wrapper based feature selection methods come into operation. Guyon et al. [100] discuss a number of wrapper methods for feature selection, however, similar to the works of Laukka et al. [387] and Bone et al. [270], we use a brute-force Sequential Forward Search (SFS) algorithm; with classification accuracy as the objective function for SFS.

6.5.5 Experimental Results and Discussion

We use the SVM classifier [116] with libSVM implementation [169] for training a model to classify between the speech of TD children and children with ASD. The dataset is divided into training and test partitions, with the classifier being trained on the training partition and tested only once on the test partition. In order to optimise hyper-parameters for SVM, the training partition is further partitioned into sub-training and sub-development partitions. Several SVM classifiers are trained with linear and RBF kernels, and a grid search is carried out to optimise model parameters for maximising classification accuracy on the sub-development partition. Parameters of the grid search are: cost $C = \{10^{-7}, 10^{-6}, \cdots 10^{3}\}$ and RBF kernel $G = \{2^{-8}, 2^{-4}, \cdots 2^{8}\}$.

Each feature in the training and development partition is normalised to the range [0,1] and the same normalisation parameters are used for the test partition as well. Naturally, since the training and test partitions contain features from different subjects there is no guarantee that features in the test partition will strictly be normalised to [0,1].

A summary of classification results is given in Table 6.1. The parameter t for Rule 1, described previously in Section 6.5.1, has been optimised for maximising classification accuracy. The settings include (a) no segmentation, (b) 1 s, (c) 500 ms, (d) 200 ms, (e) 75 ms and (f) 25 ms. These results show that not segmenting speech recordings at all actually provides the largest classification accuracy i.e. 98.96%.

It is important to recall that speech recordings for the dataset were carefully segmented from longer audio recordings of interaction between interviewer and participants, such that only speech from participants was present in those recordings. We therefore believe that not implementing speech segmentation may work best when it is ensured that the dataset only contains speech from the participant being screened. This may not work well if the dataset contains extraneous information.

We had similar observations while conducting experiments for AVEC 2017 DCC, as discussed previously in Section 4.8.2, where we observed there that retaining unvoiced regions led to smaller prediction errors in terms of RMSE, as compared to removing unvoiced regions altogether.

An alternate explanation (for higher accuracy achieved without segmentation) is based on the work of Bone et al. [270] for the ComParE 2013 Autism sub-challenge [64]. There they argued that recording environment can be a major confounding factor which can be picked up by automated screening methods for ASD which rely on speech. The Autism sub-challenge used Child Pathological Speech Database (CPSD) corpus which was first published in [388]. The flaw of this dataset is that participants from typically developing group were recorded in their school whereas participants from atypically developing group were recorded either at their homes or clinics. Bone et al. [270] demonstrated that it was possible to achieve reasonably high accuracy by building a classifier which explicitly focused on differentiating between the classes on the basis of room acoustics (see discussion provided in Section 4 of their paper [270] — thereby proving that the recording environment can indeed be a major confounding factor.

Bjorn Schuller, the main organiser of Interspeech ComParE 2013 and Fabian Ringeval, the first author of the paper introducing CPSD corpus [388] acknowledged the impact of different recordings environments on speech recordings as a significant confounding factor in their recently published paper on lessons learnt from ComParE 2013 [29].

They performed additional experiments to alleviate the influence of confounding factors which arise due to different recording environments. Their first experiment involved not using spectral features at all, since they surmised that spectral features captured differences in room acoustics. We believe this approach is not sustainable since most speech features are computed directly or indirectly through spectral analysis. As expected, they reported a significant drop in classification accuracy when all spectral features were removed.

In the second experiment, they removed all static features and used only

derivatives of feature contours. This approach to alleviate some confounding factors makes sense since dynamic features are not affected by the room acoustics, unless the room's frequency response is also time-varying. Although Schuller et al. [29] report that the classification accuracy dropped when only dynamic features were used, we believe that this is a viable approach when speech recordings are made in different environments.

Now, in light of the works of Bone et al. [270] and Schuller et al. [29], we decided to investigate features selected by SFS for the first cross-validation fold for our dataset using the two-tailed Mann-Whitney U-test [43]. While the SFS algorithm selected 29 features, we list the top 10 features in terms of their effect size in Table 6.2.

One can note that while most of the top features present are dynamic i.e. 1st or 2nd derivative of feature contours, there are some static features as well. Although, as attested by the Mann-Whitney U-test p-values and effect size, the difference between feature values are not marginal. Furthermore, to provide an illustration of the efficacy of the selected features we show the separability of TD/ASD classes after projection of 29 features from the first CV fold onto two principal components while retaining 78.40% of Eigenvalue energy in Figure 6.2. feedback from the lessons learned paper from Schuller et al. [29]

Therefore, on the basis of experiments performed in this thesis and in light of published literature, we argue that while the DSM-5 [18] does not currently recognise that speech production is affected by ASD, there is enough evidence to suggest that ASD does indeed affect it.

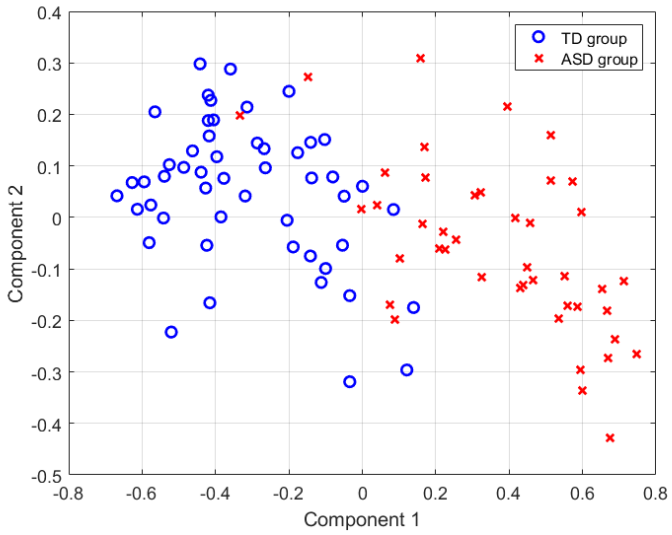


Figure 6.2: Speech features from first CV fold projected onto two principal components

Table 6.2: Mann-Whitney U-test analysis of top 10 features from our brute force classification approach

Feature Name	TD med.	ASD med.	p-value	z-value	Effect size
`Formant1_RAW_1stDeriv_StdDev`	4.90e-01	2.03e-01	1.33e-15	7.99	0.82
`Formant1_RAW_2ndDeriv_StdDev`	4.81e-01	2.05e-01	2.46e-15	7.92	0.82
`MFCC4_RAW_AbsDiffIQRStdDev`	6.18e-01	1.65e-01	2.78e-15	7.90	0.81
`Formant1_RAW_1stDeriv_AbsDiffIQRStdDev`	5.37e-01	2.63e-01	3.54e-15	7.87	0.81
`Formant1_RAW_StdDev`	4.64e-01	2.19e-01	8.23e-15	7.76	0.80
`Formant1_ZSCORE_Min`	7.67e-01	4.67e-01	2.86e-14	7.60	0.78
`Formant1_RAW_2ndDeriv_AbsDiffIQRStdDev`	5.24e-01	1.66e-01	4.73e-13	7.23	0.75
`MFCC5_RAW_InterQuartileRange`	5.61e-01	1.90e-01	6.61e-13	7.19	0.74
`MFCC4_ZSCORE_1stDeriv_Max`	2.45e-01	5.44e-01	7.81e-13	-7.16	-0.74
`MFCC4_ZSCORE_Max`	2.60e-01	5.57e-01	2.01e-14	-7.65	-0.79

6.6 Statistical Analysis for Standard Audio Feature Sets

In addition to the discussion on our proposed mechanism for feature engineering and classification for automated screening of ASD from speech, we have also performed statistical analysis of features defined in three standard feature sets i.e. the OpenSmile Prosody, the OpenSmile eGeMAPS, and the COVAREP Voice Quality. Our motivation for undertaking statistical analysis of standard feature sets are as follows:

- The primary motivation for using standard feature sets is that these can enable comparison between datasets which are private i.e. which cannot be shared due to ethical restrictions.

  While we acknowledge the existence of these restrictions, we argue that discriminative ability of features can still be shared publicly, thereby enabling progress in research even across private datasets.

- Our secondary motivation is to investigate the efficacy of voice quality features for discrimination between speech of individuals from TD and ASD groups. We aim to compare traditional voice quality features such as shimmer, jitter, and harmonics-to-noise ratio [46] against more recently developed voice quality features which are available with the COVAREP feature set [62].

We use the Mann-Whitney U-test for statistical analysis of features, inline with the works of [145, 389]. The U-test offers a non-parametric means to test the null hypothesis that it is equally likely that a randomly selected feature from one group (say TD group) will be less than or greater than a randomly selected feature from the second group (ASD group).

We thus test the null hypothesis that features from the TD and ASD group are samples from continuous distributions with equal medians, against the alternative hypothesis that they are not. We set a p-value of 0.05 as the cut-off for significance for feature analysis with the U-test. If the p-value is indeed smaller than the cut-off value, we reject the null hypothesis and conclude that a significant difference does exist between median values of the two features, meaning that the feature has ability to discriminate between TD and ASD groups. Inspired by statistical analysis conducted by [28], we do not make any prior hypothesis on the effectiveness of a particular feature and therefore use *two-tailed* U-test for hypothesis testing instead of a single-tailed test. Finally, we compute the effect size using the z-statistic of the U-test based on the work of Fritz et al. [45]. Thus, in addition to the p-values for the U-test, we also provide the magnitude of the difference between groups.

6.6.1 Pre-processing

We discussed in Section 6.5.5 that speech segmentation has a significant effect on the classification accuracy. Therefore, one cannot arbitrarily compute features from the standard feature sets for our speech recording. We surmise that one needs to select a segmentation duration so that other researchers have detailed guidelines on recreating experiments.

We therefore propose the following set of pre-processing steps.

- A speech segment should have at least 150 ms of voiced speech. This is roughly based on the requirement for Summation of the Residual Harmonics (SRH) based pitch tracker [380] available with the COVAREP toolbox [31] i.e. to have at least 100 ms worth of speech recording. By fixing a requirement to have 150 ms or more of voiced speech, our aim is to prioritise computing features from voiced speech rather than unvoiced speech. This is because unvoiced speech as detected from voice activity detectors can also mean silence [382].

- Unvoiced speech up to a duration of 300 ms needs to be tolerated. In case unvoiced speech is greater than 300 ms, then any duration greater than 300 ms is removed and the speech signal is stitched back together. We select 300 ms duration since it is used for the DIAC-WOZ corpus which was used for depression recognition sub-challenges for AVEC 2016 DCC [12] and AVEC 2017 DSC [13]. These settings appear to have been accepted by the research community at large.

- We use voice activity detector (VAD) [381] from Drugman et al. for identifying voiced and unvoiced speech in speech recordings. Drugman et al. provide open source implementation for this code which is available with COVAREP toolbox. The authors also integrate features from prior work of Sadjadi et al. [390] to further increase robustness of their VAD. We fix 0.5 as the threshold for voice probability to differentiate between voiced and unvoiced speech.

- Finally, we use min-max normalisation to ensure that the dynamic range of the speech signal lies within [-1,1] prior to computing features.

6.6.2 OpenSmile Prosody Feature Set

Abnormal and disorder prosody is an established clinical marker for communication deficits which occur due to ASD [182, 369, 391]. Given this, we investigate the discriminative power of pitch and loudness; two features commonly used to characterise speech prosody [46].

To this end, we use the OpenSmile Prosody feature [30] to compute the pitch (in both linear and log scale) and loudness. For these features, a set of seven functionals are computed, which include mean, standard deviation,

Table 6.3: Mann-Whitney U-test analysis of prosody features with pitch range 52–622 Hz (default setting)

Feature Name	**TD med.**	**ASD med.**	p	z	r
		BEDROOM			
`logPitch (range)`	6.15e+00	3.76e+00	1.85e-03	3.11	0.45
`Pitch (range)`	8.84e+01	4.86e+01	5.05e-03	2.80	0.40
`logPitch (stddev)`	2.76e+00	2.05e+00	1.19e-02	2.51	0.36
`Pitch (stddev)`	4.22e+01	2.65e+01	1.88e-02	2.35	0.34
`Loudness (slp)`	-6.41e-06	-2.43e-05	1.88e-02	2.35	0.34
`logPitch (Perc. 90)`	4.17e+01	3.90e+01	2.48e-02	2.25	0.32
`Pitch (Perc. 90)`	3.06e+02	2.62e+02	2.90e-02	2.18	0.32
`Pitch (mean)`	2.60e+02	2.33e+02	3.57e-02	2.10	0.30
`logPitch (mean)`	3.86e+01	3.68e+01	4.15e-02	2.04	0.29
`Loudness (stddev)`	5.53e-01	4.85e-01	4.37e-02	2.02	0.29
`Loudness (Perc. 10)`	4.03e-01	6.22e-01	1.49e-03	-3.18	-0.46
		CARTOON			
`logPitch (range)`	6.72e+00	3.13e+00	4.20e-05	4.10	0.62
`Pitch (range)`	1.03e+02	4.27e+01	2.56e-04	3.66	0.56
`Pitch (stddev)`	4.82e+01	2.63e+01	1.00e-03	3.29	0.50
`Pitch (Perc. 90)`	3.10e+02	2.57e+02	1.00e-03	3.29	0.50
`logPitch (Perc. 90)`	4.19e+01	3.87e+01	1.10e-03	3.26	0.50
`logPitch (stddev)`	3.09e+00	1.95e+00	1.30e-03	3.22	0.49
`Loudness (stddev)`	6.35e-01	4.63e-01	6.88e-03	2.70	0.41
`Loudness (range)`	1.62e+00	1.19e+00	1.50e-02	2.43	0.37
`Pitch (mean)`	2.60e+02	2.30e+02	3.44e-02	2.12	0.32
`logPitch (mean)`	3.87e+01	3.66e+01	4.90e-02	1.97	0.30
`Loudness (Perc. 10)`	4.51e-01	5.45e-01	7.41e-03	-2.68	-0.41

the 10th percentile, the 90th percentile, and the outlier robust range which is computed as the difference between 99th percentile and the 1st percentile.

We start our work by first investigating the influence of pitch ceiling on the discriminative ability of prosody features. The basis of this investigation is the work of Kiss et al. [276] who report that pitch ceiling can influence the discriminative power of pitch based features. To this end, we compute prosody features for two pitch ranges: (a) 52–622 Hz i.e. the default setting for the OpenSmile Prosody feature set and (b) 50–700 Hz i.e. the range typically used for speech of children [46]. Results in terms of Mann-Whitney U-test for this investigation are summarised in Table 6.3 for the pitch range of 52–622 Hz and in Table 6.4 for the pitch range 50–700 Hz. Here we report the median values for TD and ASD groups, the p-value of the Mann-Whitney U-test, its z-value,

Table 6.4: Mann-Whitney U-test analysis of prosody features with pitch range 50–700 Hz

Feature Name	**TD med.**	**ASD med.**	p	z	r
		BEDROOM			
`logPitch (Range)`	6.15e+0	3.76e+0	1.72e-3	3.13	0.45
`Pitch (Range)`	8.86e+1	4.94e+1	4.74e-3	2.82	0.41
`logPitch (stddev)`	2.77e+0	2.05e+0	1.42e-2	2.45	0.35
`loudness (slp)`	-6.41e-6	-2.43e-5	1.88e-2	2.35	0.34
`Pitch (stddev)`	4.23e+1	2.66e+1	2.22e-2	2.29	0.33
`logPitch (Perc. 90)`	4.17e+1	3.90e+1	2.35e-2	2.27	0.33
`Pitch (Perc. 90)`	3.06e+2	2.62e+2	2.61e-2	2.22	0.32
`Pitch(mean)`	2.60e+2	2.33e+2	3.57e-2	2.10	0.30
`logPitch(mean)`	3.86e+1	3.68e+1	3.95e-2	2.06	0.30
`Loudness (stddev)`	5.53e-1	4.85e-1	4.37e-2	2.02	0.29
`Loudness (Perc. 10)`	4.03e-1	6.22e-1	1.49e-3	-3.18	-0.46
		CARTOON			
`logPitch (Range)`	6.70e+0	3.13e+0	4.20e-5	4.10	0.62
`Pitch (Range)`	1.03e+2	4.31e+1	2.56e-4	3.66	0.56
`Pitch (stddev)`	4.99e+1	3.14e+1	8.43e-4	3.34	0.51
`logPitch (stddev)`	3.09e+0	2.12e+0	1.00e-3	3.29	0.50
`Pitch (Perc. 90)`	3.11e+2	2.58e+2	1.10e-3	3.26	0.50
`logPitch (Perc. 90)`	4.19e+1	3.87e+1	1.19e-3	3.24	0.49
`Loudness (stddev)`	6.35e-1	4.63e-1	6.88e-3	2.70	0.41
`Loudness (Range)`	1.62e+0	1.19e+0	1.50e-2	2.43	0.37
`Pitch( mean)`	2.60e+2	2.31e+2	3.44e-2	2.12	0.32
`logPitch (mean)`	3.87e+1	3.66e+1	4.62e-2	1.99	0.30
`Loudness (Perc. 10)`	4.51e-1	5.45e-1	7.41e-3	-2.68	-0.41

and the effect size r. A comparison of these tables reveals that pitch ceiling only has a minor affect on the discriminator power of pitch based features, although one can argue that the default pitch range of Prosody features i.e. 622 Hz is close to the pitch range for children (700 Hz) therefore the influence of pitch ceiling is small.

Nevertheless, based on Fusaroli et al. [46], we continue our investigation of prosody features using results from the pitch ceiling of 50–700 Hz, i.e. results as summarised in Table 6.4. Here we observe that pitch and loudness variability is generally larger for the TD group as compared to the ASD group. Moreover, the average pitch is also greater for subjects from the TD group as compared to subjects from the ASD group. This reinforces the point of view that individuals with ASD have diminished affect in their speech communication,

with a perceptually monotonic speech. These results are in line with the works of DePape et al. [182] and Kaland et al. [392] who also reported similar observations.

We also report that we do not find a difference between the discriminative power of linear scale and log scale pitch features, even though Bone et al. [195] preferred log scale pitch features based on the understanding that log scale pitch features were more perceptually relevant. Finally, we report that we did not observe difference in observations for Prosody features between BEDROOM and CARTOON experiments.

6.6.3 OpenSmile eGeMAPS Feature Set

We now investigate the efficacy of features from the extended Geneva minimalistic acoustic parameter set (eGeMAPS) feature set [58]. The eGeMAPS feature set contains an expert knowledge based minimalistic set of features which provide information about prosody, voice quality, formants, and spectral characteristics of speech. These features have proven to be effective in recognising changes in voice production due to emotional affect and physiology [58]. Furthermore, in recent past, Pokorny et al. [145] had also carried out Mann-Whitney U-test based statistical analysis for eGeMAPS feature set as part of their work on screening for ASD from speech. Therefore it makes sense to investigate the efficacy of these features for our dataset as well.

We compute eGeMAPS feature set using the OpenSmile toolkit [30]. This feature set contains 88 features which are based on functionals of acoustic LLDs. A summary of these features is provided in Table 6.5 as a reference, although we strongly encourage the reader to refer the tutorial on eGeMAPS feature set by Eyben et al. [58].

The results of eGeMAPS features computed for BEDROOM and CARTOON experiments are shown in Tables 6.6 and 6.7, respectively. The observation that TD participants have larger pitch and loudness variability is found true from eGeMAPS features as well.

Interestingly, we do not find any of the traditional voice quality features such as shimmer, jitter, and HNR to be discriminative beyond the p-value threshold. This means that those voice quality features are not useful for discriminating between speech of individuals with ASD. However, our investigation into the affect of ASD on voice quality is not complete, and will continue in the next section.

Meanwhile, we find several spectral features to have high discriminative power. For example, the `slpUV500-1500 (mean)` feature which provides information about the slope for unvoiced speech in the spectral range of 500-1500 Hz is not only statistically significant but also has an effect size of 0.73 for the BEDROOM experiment and an effect size of 0.60 for the CARTOON experiment. Conversely, we note that the feature `slpUV500-1500 (stddev)`

has an effect size of -0.57 for the BEDROOM experiment and an effect size of -0.65 for the CARTOON experiment.

These results are intriguing because there is a chance that these features, based on unvoiced speech, may have picked up the recording environment, as opposed to speech impairment due to ASD. It is reminded that like other datasets with speech recording of TD and ASD groups [64, 145, 366, 388], our dataset also includes somewhat different recording environments. While it causes an undesired bias in datasets, there are usually ethical restrictions which force that subjects, especially those with special needs, be recorded in environments where they feel most comfortable.

A counter argument to potential bias in the dataset is that spectral characteristics of unvoiced speech is indeed a valid difference between the two groups. In a recently published paper, Pokorny et al. [145] used eGeMAPS features and reported unvoiced spectral slp to be the most discriminatory in terms of effect size of Mann-Whitney tests — similar to our finding. They explicitly mention that there was no background noise in their speech recordings, although since their dataset is also private so one cannot verify their claim independently. Furthermore, in their work on exploring a small set of robust features for emotion recognition from speech, Tahon et al. [393] also found unvoiced speech features to be more useful than voiced speech features. Nevertheless, we believe that it is too soon to accept or reject the efficacy of features computed from unvoiced speech at this moment due to the scarcity of publicly available datasets and the fact that only a few researchers working on ASD have provided results for eGeMAPS features.

Finally, we observe that several features based on formant frequencies also register as statistically significant. This result is in line with the work of Lyakso et al. [389] who also found formant frequencies to be useful for screening individuals with ASD using speech.

6.6.4 COVAREP Voice Quality Feature Set

The Cooperative Voice Analysis Repository for Speech Technologies (COVAREP) feature set is a standard feature set which was introduced by Drugman et al. [31] in 2014, with the most recent update taking place in May 2018.

As discussed in Section 3.4.2, the COVAREP toolbox provides a number of new voice quality features which include normalised amplitude quotient (NAQ), quasi-open quotient (QOQ), parabolic spectral parameter (PSP), maxima dispersion quotient (MDQ), harmonic-to-noise ratio (HNR), corrected difference of first two harmonic amplitudes (H1H2), cepstral peak prominence (CPP), and peak slope (PS). These voice features have shown promise, especially for the task of automated screening of depression [125, 126, 394]. Since the COVAREP toolbox computes acoustic LLDs only, we use mean, standard deviation, and range functionals to create a global representation of the feature for the entire recording.

Table 6.5: Summary of features in the eGeMAPS feature set

Features	**Functionals**	**V/UV/VUV/–**
	Prosody	
Pitch	mean, stddev, Perc20, Perc50 Perc80, range(Perc20,Perc80) mean of rising/falling slp stddev of rising/falling slp	V
Loudness	mean, stddev, Perc20, Perc50 Perc80, range(Perc20,Perc80) mean of rising/falling slp stddev of rising/falling slp	VUV
Loudness peaks per sec	–	–
Voiced segs. per sec	–	–
Voiced segs. length	mean & stddev	–
UnVoiced segs. length	mean & stddev	–
Equivalent sound level	–	–
	Voice Quality	
Jitter	mean & stddev	V
Shimmer	mean & stddev	V
HNR	mean & stddev	V
	Spectral	
Alpha ratio	mean & stddev	V & U
Hammarberg index	mean & stddev	V & U
Slope slp0-500	mean & stddev	V & U
Slope slp500-1500	mean & stddev	V & U
H1-H2	mean & stddev	V
H1-A3	mean & stddev	V
MFCCs 1-4	mean & stddev	V & U
Spec. flux	mean & stddev	V & U
	Formants	
Formants 1-3	mean & stddev	V
Formants 1-3 BW	mean & stddev	V
Formants 1-3 rel. energy	mean & stddev	V

Table 6.6: Mann-Whitney U-test analysis of eGeMAPS features for BEDROOM experiment

Feature Name	**TD med.**	**ASD med.**	p	z	r
`slpUV500-1500 (mean)`	-8.79e-3	-1.47e-2	3.78e-7	5.08	0.73
`Pitch (range)`	3.70e+0	2.32e+0	4.19e-4	3.53	0.51
`SpecFluxV (stddev)`	6.28e-1	5.28e-1	8.33e-4	3.34	0.48
`Loudness (stddev)`	5.25e-1	4.35e-1	1.85e-3	3.11	0.45
`SpecFlux (stddev)`	7.88e-1	6.52e-1	2.79e-3	2.99	0.43
`slpV500-1500 (mean)`	-2.19e-2	-2.51e-2	3.19e-3	2.95	0.43
`MFCC4 (stddev)`	3.12e-1	2.17e-1	5.74e-3	2.76	0.40
`H1-H2 (mean)`	7.14e+0	3.82e+0	7.37e-3	2.68	0.39
`MFCC1 (stddev)`	5.01e-1	4.37e-1	1.59e-2	2.41	0.35
`Pitch (Perc. 80)`	4.06e+1	3.78e+1	1.78e-2	2.37	0.34
`F1 (stddev)`	2.52e-1	2.29e-1	1.88e-2	2.35	0.34
`MFCC3 (stddev)`	2.42e+0	1.22e+0	1.99e-2	2.33	0.34
`Pitch (stddev)`	7.61e-2	5.73e-2	2.22e-2	2.29	0.33
`MFCC3 (mean)`	4.98e+0	2.94e+0	2.75e-2	2.20	0.32
`Pitch (mean)`	3.86e+1	3.67e+1	4.59e-2	2.00	0.29
`H1-H2 (stddev)`	1.45e+0	2.26e+0	4.37e-2	-2.02	-0.29
`slpUV0-500 (mean)`	-3.36e-2	-2.24e-2	4.15e-2	-2.04	-0.29
`Loudness (Perc. 20)`	5.58e-1	6.97e-1	2.75e-2	-2.20	-0.32
`MFCC4 (stddev)`	-1.14e+0	-8.93e-1	1.27e-2	-2.49	-0.36
`F1BW (mean)`	1.35e+3	1.40e+3	1.27e-2	-2.49	-0.36
`F2BW (mean)`	1.13e+3	1.21e+3	5.29e-4	-3.47	-0.50
`slpV500-1500 (stddev)`	-6.56e-1	-5.01e-1	8.82e-5	-3.92	-0.57

Similar to Prosody features, we first investigated the influence of pitch ceiling on the discriminative power voice quality features. To this end, we compared results for features computed with the default ceiling of the COVAREP toolkit i.e. 500 Hz with features computed with a pitch ceiling of 700 Hz. Our results indicated there was no effect on voice quality features due to difference in pitch ceiling. However, in continuation of our approach used for Prosody, we shall report results for features computed with a pitch ceiling of 700 Hz, as summarised in Table 6.8.

An interesting observation from Table 6.8 is that several voice quality features from the COVAREP toolkit register as statistically significant even though none of the traditional voice quality features registered as statistically significant during our statistical analysis of the eGeMAPS dataset.

Amongst these, one finds that feature values for `MDQ_stddev`, `H1H2 (stddev)`, `peakslp (stddev)` are larger for the TD group whereas `Rd_conf (range)` and `Rd_conf (stddev)` are larger for the ASD group. These results demonstrate

Table 6.7: Mann-Whitney U-test analysis of eGeMAPS features for CARTOON experiment

Feature Name	**TD med.**	**ASD med.**	p	z	r
`Pitch (range)`	3.94e+0	1.94e+0	1.13e-5	4.39	0.67
`slpUV500-1500 (mean)`	-8.08e-3	-1.47e-2	8.67e-5	3.93	0.60
`slpV500-1500 (mean)`	-2.06e-2	-2.57e-2	5.39e-4	3.46	0.53
`Loudness (stddev)`	5.38e-1	4.35e-1	5.90e-4	3.44	0.52
`SpecFluxV (stddev)`	6.24e-1	5.24e-1	1.10e-3	3.26	0.50
`SpecFlux (stddev)`	7.97e-1	6.67e-1	1.67e-3	3.14	0.48
`MFCC1V (stddev)`	3.53e-1	2.15e-1	1.67e-3	3.14	0.48
`Pitch (stddev)`	8.21e-2	5.83e-2	3.47e-3	2.92	0.45
`Loudness (stddev slp)`	7.18e+0	5.54e+0	5.11e-3	2.80	0.43
`Pitch (Perc. 80)`	4.04e+1	3.75e+1	9.20e-3	2.60	0.40
`MFCC1 (stddev)`	4.65e-1	4.09e-1	2.69e-2	2.21	0.34
`Loudness (range)`	1.04e+0	8.23e-1	2.86e-2	2.19	0.33
`F1BW (mean)`	1.34e+3	1.36e+3	3.88e-2	-2.07	-0.32
`F2BW (mean)`	1.10e+3	1.17e+3	3.73e-4	-3.56	-0.54
`slpV500-1500 (stddev)`	-6.21e-1	-4.59e-1	2.20e-5	-4.24	-0.65

the promise of voice quality features from the COVAREP toolkit for the task of discriminating between speech of individuals from TD and ASD groups. Thus, through this work, we propose, for the very first time, the use of these features for the task at hand.

There is, however, a caveat to the use of COVAREP feature set. In line with our discussion in Section 3.4.2, we believe one cannot use voice quality features to discriminate speech on a perceptual scale between breathy to tense unless one has the value of these features for modal voice of each subject, even if these features can discriminate between TD and ASD groups. This is because the amplitude of these features is known to be subject dependent [234, 241], which means that with a reference for modal voice, one cannot conclude that individuals with TD have a breathier or tense voice compared to the ASD group.

6.7 Summary

In this chapter we discussed our work towards the development of automated methods for screening Autism Spectrum Disorder. The contributions of our work are listed as follows:

- We manually annotated audio/visual recordings of interview sessions using ELAN software [41, 42] to mark segments of recordings which only

Table 6.8: Mann-Whitney U-test analysis of Voice Quality features

Feature Name	**TD med.**	**ASD med.**	p	z	r
		BEDROOM			
`MDQ (range)`	2.07e-1	1.72e-1	1.47e-4	3.80	0.55
`MDQ (stddev)`	2.74e-2	2.37e-2	3.31e-4	3.59	0.52
`Ps (stddev)`	6.50e-2	5.37e-2	6.16e-4	3.42	0.49
`H1-H2 (stddev)`	6.45e+0	5.52e+0	1.12e-3	3.26	0.47
`PSP (range)`	2.47e+0	1.68e+0	2.12e-3	3.07	0.44
`PSP (stddev)`	3.48e-1	1.75e-1	2.43e-3	3.03	0.44
`Ps (range)`	3.87e-1	3.22e-1	3.41e-3	2.93	0.42
`MDQ (mean)`	1.30e-1	1.36e-1	3.95e-2	-2.06	-0.30
`Rd (stddev)`	7.68e-2	8.46e-2	2.90e-2	-2.18	-0.32
`NAQ (mean)`	8.65e-2	1.31e-1	1.13e-2	-2.53	-0.37
`QOQ (range)`	7.36e-1	8.35e-1	8.34e-3	-2.64	-0.38
`QOQ (mean)`	2.55e-1	4.40e-1	2.79e-3	-2.99	-0.43
		CARTOON			
`Ps (stddev)`	6.67e-2	4.71e-02	2.82e-04	3.63	0.55
`H1-H2 (stddev)`	6.40e+0	5.40e+00	1.42e-03	3.19	0.49
`Ps (range)`	4.02e-1	3.52e-01	1.06e-02	2.56	0.39
`Rd (stddev)`	7.73e-2	8.59e-02	3.04e-02	-2.16	-0.33
`Rd (range)`	5.42e-1	6.01e-01	2.22e-02	-2.29	-0.35

contain speech of subjects. This enabled us to undertake investigation into automated screening for ASD using speech.

While our current work is limited to analysis of speech, our efforts for annotation of these recordings opens up avenues for future research. For example, analysis of facial expressions and body movement when children either speak or are spoken to by the interviewer. Using these annotations, models can also be built to investigate synchrony of dyadic communication between children and interviewer.

We acknowledge and credit Dr. Catherine Jones (from Cardiff University's School of Psychology) and her team for collecting and providing us these audio/visual recordings. Our work focuses explicitly on social signal processing, not data collection.

- We discussed our feature engineering and classification mechanism for automated screening of ASD from speech for subjects in our dataset. The proposed feature engineering mechanism consists of a four step process which includes speech segmentation, feature extraction, feature description, and feature selection.

The efficacy of this mechanism was demonstrated in terms of classification accuracy and the identification of highly discriminative speech features using Mann-Whitney U-test based statistical analysis [43] and effect size for statistically significant features [44, 45].

- We investigated the influence of speech segmentation on classification accuracy. To this end we performed experiments using six different segmentation rules. Our results suggest that classification accuracy is dependent on the duration of voiced speech in each speech segment.

- We performed several experiments to identify discriminative features defined in three standard feature sets i.e. the COVAREP [31], openSmile Prosody [30], and openSmile eGeMAPS [58].

  Our motivation stems from the fact that most researchers working towards the development of automated screening methods for ASD cannot share their dataset due to ethical restrictions. While we acknowledge the existence of such restrictions (we cannot share our dataset either), we argue that discriminative power of features in standard feature sets can be reported openly even if the dataset itself cannot be shared due to ethical restrictions. This will enable progress in research for the development of automated screening methods even across private datasets.

- We report from experiments conducted using the standard feature sets that subjects with ASD have smaller pitch and loudness variability (in terms of standard deviation), which suggests monotonic speech. This is in line with findings in [46].

- We report on the basis of our experiments that traditional voice quality features such as shimmer, jitter, and HNR are not able to provide discrimination between speech of individuals from TD and ASD groups. In addition, we report that features from the COVAREP voice quality feature set are able to discriminative between the speech of individuals from the two groups.

- Finally, on the basis of experiments performed in this chapter and in light of published literature, we argue that while the DSM-5 [18] does not currently recognise that the speech production system is affected by ASD, there is enough evidence from research literature as well as our investigation to suggest that ASD may actually have a significant effect on the vocal production system.

Chapter 7

Conclusion and Future Work

7.1 Introduction

In this thesis, the development of automated screening methods for mental and neuro-developmental disorders was discussed. To this end, we proposed methods for automated screening of depression, bipolar disorder, and autism spectrum disorder (ASD) based on audio/visual modalities.

Mental and neuro-developmental disorders are critical health issues which affect a large number of people. Depression, according to the World Health Organisation (WHO), is the largest cause of disability worldwide and affects more than 300 million people [278]. Bipolar disorder affects more than 60 million individuals worldwide [325]. ASD, meanwhile, affects more 1 in 100 people in the UK [357]. Not only do these disorders affect individuals personally, they also have a significant economic impact.

Clinicians have long reported that early identification of these disorders followed by suitable treatment can help individuals lead a relatively normal life or at least improve their quality of life. This has motivated computer scientists to explore possibilities of developing automated methods to screen for these disorders. In recent years, there has been a plethora of work towards development of computational methods for this purpose. However, one finds that a number of publications focus on brute-force approaches where the objective is to maximise accuracy for a particular dataset. This is a commonly followed path in machine learning.

While these approaches are potentially useful for learning new features which could be representative of these disorders, such approaches may not be best suited for developing robust computational methods for screening of these disorders [266]. This is due to a myriad of confounding factors including human factors, which affect symptoms of these disorders.

The main objective of this thesis was to develop, investigate, and propose computational methods, in particular features and machine learning pipelines, which capture the traits of these disorders in accordance with descriptions

given in the Diagnostic and Statistical Manual (DSM-5) [18]. The DSM-5 manual is a guidebook published by the American Psychiatric Association which offers common language on mental disorders. Our motivation was to alleviate, to an extent, the possibility of machine learning algorithms picking up one of the confounding factors to optimise performance for the dataset — something which we do not find uncommon in research literature.

For the task of automated screening of depression, we demonstrated the effectiveness of our proposed approaches on two publicly available datasets [11, 12], as well as the depression recognition challenge at the 2017 edition of the Audio/Visual Emotion recognition Challenge (AVEC) workshop [13]. Similarly, for automated screening of bipolar disorder, we demonstrated the efficacy of our proposed approaches for the 2018 edition of the AVEC workshop [39]. Meanwhile, we found that the task of automated screening for ASD is much more complicated. Here, confounding factors can overwhelm socials signals which are effected by ASD. We discuss, in light of research literature, that significant collaborative work is required between computer scientists and clinicians to discern social signals which are robust to common confounding factors. Finally, we discussed Cardiff University's proposed approaches for 2017 and 2018 editions of computational paralinguistics challenges (ComParE) [146, 148], as well as lessons learned from these challenges.

7.2 Summary of Thesis Achievements

- We participated and proposed solutions for four major competitions in SSP/AC, namely the AVEC 2017 Depression severity prediction sub-challenge, the AVEC 2018 Bipolar disorder sub-challenge, the ComParE 2017 Cold sub-challenge, and ComParE 2018 challenges. As a result two of our papers were published and one has been accepted for publication.

- We surmised that individuals with mental disorders, such as depression and bipolar disorder, have uniqueness to their facial muscle movement and speech which manifest as sudden and erratic changes to contours of audio/visual features. To this end, we proposed a novel set of temporal features, which we call *turbulence features*, to quantify fluctuations in contours of these features.

  We initially proposed turbulence features for predicting depression severity as part of our solution to the AVEC 2017 Depression severity prediction sub-challenge [13]. We beat the challenge baselines whilst using turbulence features and stood sixth overall.

  Turbulence features were also used as part of our solution for the AVEC 2018 Bipolar disorder sub-challenge. While we did not beat the challenge baseline, we managed to exactly match it. The overall standings for AVEC

2018 challenges will be revealed at the ACM Multimedia Conference to be held in October 2018.

- We detailed a methodology to quantify specific craniofacial movements, which we hypothesised could be indicative of psychomotor retardation and hence depression. The efficacy of these features was tested in terms of Pearson's correlation coefficient with respect to depression severity. In order to demonstrate the efficacy of these features across datasets, we used three sets of recordings from two publicly available datasets from AVEC challenges on depression severity prediction i.e. AVEC 2014 Depression Severity prediction Challenge (DSC) [11] and AVEC 2017 DSC [13]. Given that these features are inspired by knowledge of psychomotor retardation from the DSM 5 manual [18], we believe that interpretability of these features will provide meaningful feedback to clinicians for diagnosis of depression.

- We hypothesised that individuals with depression have unique characteristics to their speech spectra. To this end, we introduced Fisher vector encoding [36, 99] of spectral low level descriptors (LLDs) for quantifying abnormalities within speech spectra of individuals with depression.

  Initially, we demonstrated the efficacy of our proposed approach for the AVEC 2016 Depression classification sub-challenge (DCC) dataset [12], where the objective was to identify individuals with and without depression [37]. Later, we extended the idea by adding temporally-piecewise aggregation of Fisher vectors as part of our solution to the AVEC 2017 DSC [34]. We beat the challenge baseline whilst using this method.

- We introduced Fisher vector encoding of Computational Paralinguistics Challenge (ComParE) low level descriptors and demonstrate that these features are viable for predicting the severity of mania.

  Our motivation for using ComParE LLDs is based on the fact that these LLDs contain features which represent information pertaining to speech prosody, voice quality, and speech spectra [30]. Therefore, intuitively, by using these LLDs, Fisher Vector features should represent even more detailed information about the characteristics of speech for individuals with bipolar disorder, compared to spectral features alone.

  We also show that these features perform much better than ComParE functionals [64] for the AVEC 2018 Bipolar disorder sub-challenge.

- We introduced the Greedy Ensembles of Weighted Extreme Learning Machines (GEWELMs) classifier as part of our solution for the Interspeech ComParE 2018 challenge, which involved multi-class classification tasks with a large imbalance of label distribution. GEWELMs combine the well-known training efficiency of Extreme Learning Machines (ELM) [123]

with good classification performance. This combination of speed and accuracy, we speculate, will be especially important in real-time scenarios, such as automated screening.

We demonstrated the efficacy of GEWELMs not only for ComParE 2018 challenges but also for the AVEC 2018 Bipolar disorder sub-challenge. Moreover, we compared GEWELMs with implementations of linear SVM [116] based on libSVM [169] and LIBLINEAR toolkits [171], and found GEWELMs to perform better than linear SVM.

- As part of our work on automated screening methods for Autism Spectrum Disorder (ASD), we conduct a number of experiments and provide discussion on a previously unpublished dataset.

  We started with raw audio/visual recordings of children with ASD, provided by Dr. Catherine Jones from Cardiff School of Psychology, and performed manual annotation for speech belonging to children. This enabled us to perform various experiments to investigate the effect of speech segmentation regimes on the accuracy with which automated screening methods can differentiate between speech of subjects from typically developing (TD) and ASD groups. We report the most discriminative features from our proposed feature engineering and classification mechanism. In addition to this, we report on most discriminative features from three audio feature sets which are standard in the field of SSP/AC. Inspired from the work of Kiss et al. [276], we also report on the effects of pitch ceiling and the choice of software on the discriminative power of speech based features.

  Finally, on the basis of experiments performed as part of our on automated screening of ASD and in light of published literature, we argue that while the DSM-5 [18] does not currently recognise that the speech production system is affected by ASD, there is enough evidence from research literature as well as our investigation to suggest that ASD may actually have a significant effect on the vocal production system.

7.3 Limitations

Automated screening for mental and neuro-developmental disorders is still very much a work in progress. There are a number of limitations which need to be acknowledged and challenges which need to be overcome in order to bring such systems into the public sphere. As discussed in Section 3.5, these limitations arise due to the availability of datasets, the size of these datasets, how ground truth labels are created, and the existence of confounding factors which affect the quality of data.

Arguably the biggest limitation of our work is the relatively small size of datasets which were available to us. We used datasets from AVEC 2014 [11],

2016 [12], 2017 [13], and 2018 [39] challenges in order to develop methods for automated screening of depression and bipolar disorder. Amongst these, the AVEC 2014 dataset contains 300 video recordings, the AVEC 2016/2017 datasets contains 189 video recordings, and the AVEC 2018 contains 218 video recordings. While these datasets form the largest set of publicly available multimedia datasets which can be used for the development of automated screening methods, the size of these datasets is relatively small especially in light of the difficult task of determining the mental state of individuals. This is, however, an inherent limitation of the field where data collection and sharing is restricted due to privacy concerns.

The second limitation of our work is associated with how ground truth labels are defined in these datasets. While automated screening systems can indeed bring objectivity to the screening process, these systems are still trained on data which has inherently subjective labels. This limitation, however, comes directly from the field of psychology, since mental and neuro-developmental disorders are at best diagnosed based on behavioural symptoms [28, 33, 264]. In fact, for the AVEC 2017 challenge on depression severity prediction, we note that certain participants did not complete the self-assessment questionnaires truthfully which means that quality of ground truth labels has been compromised (see Section 3.5.4 for details). Nevertheless, as pointed out by Solomon et al. [28], despite the inherent flaws of self-assessment forms, they provide at least a reasonable and quantifiable standard to measure depression against in the absence of physical tests to detect depression.

Overall, in light of discussion in this thesis, we believe that development of methods for automated screening of mental and neurological disorders can greatly benefit from collaboration between researchers from SSP/AC community and clinicians. Active collaboration can enable informed research progress which can avoid misleading deliverables, even in the presence of inherent limitations.

7.4 Future Work

We believe that research conducted as part of this thesis provides us an excellent opportunity to continue contributing towards the development of automated screening methods. Following are some pathways to extend our work.

Automated Screening of Depression

In this thesis our focus was directed towards the development of automated screening methods from audio/visual modalities. However, recent advances in natural language processing (NLP) [93, 94] mean that the text modality has a very important role to play. Furthermore, speech-to-text application programming interfaces (API) provided by vendors such as Google, IBM, and Microsoft provide us the opportunity to automatically generate transcripts

for speech recordings. Accordingly, we believe that using speech-to-text APIs followed by text analysis is a possible extension of our work, especially given that one can combine results from multiple modalities to develop even better screening methods.

We would also like to collect new data for individuals with depression from different cultural backgrounds and language. To this end, we aim to formalise a plan to collect from psychiatric hospitals in Pakistan in near future.

Automated Screening of Bipolar Disorder

We would like to use speech-to-text APIs to generated transcripts for speech uttered by subjects in AVEC 2018 Bipolar disorder challenge dataset [33] and investigate the efficacy of the text modality for screening of bipolar disorder, especially in context of the DSM-5 manual [18] and the Young Mania Rating Scale [22]. A possible limitation in this regard is that labels for the test partition are not available outside the official AVEC challenge, therefore, we may have to work with a smaller sized dataset.

Automated Screening of Autism Spectrum Disorder

In continuation of our work on automated screening of ASD from speech, we aim to investigate effects of ASD on facial expressions and head movement for our dataset. Recently, some researchers have started work in this regard [69,70], therefore, it will be interesting to compare results from our dataset with theirs.

Bibliography

[1] P. Ekman and W. V. Friesen, *The Facial Action Coding System.*, 1978.

[2] P. Ekman, W. V. Friesen, and J. C. Hager, *Facial Action Coding System. Manual and Investigator's Guide.* Research Nexus, 2002.

[3] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social Signal Processing: Survey of an emerging domain," *Image and Vision Computing*, Vol. 27, No. 12, pp. 1743–1759, 2009.

[4] A. Vinciarelli, M. Pantic, H. Bourlard, and A. Pentland, "Social Signals, Their Function, and Automatic Analysis: A Survey," in *ACM International Conference on Multimodal Interfaces (ICMI)*, 2008, pp. 61–68.

[5] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schroder, "Bridging the gap between social animal and unsocial machine: A survey of social signal processing," *IEEE Transactions on Affective Computing*, Vol. 3, No. 1, pp. 69–87, 2012.

[6] P. M. Brunet, G. McKeown, R. Cowie, H. Donnan, and E. Douglas-Cowie, "Social signal processing: What are the relevant variables? And in what ways do they relate?" in *IEEE International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2009, pp. 1–6.

[7] P. Ekman, "An argument for basic emotions," *Cognition & Emotion*, Vol. 6, No. 3, pp. 169–200, 1992.

[8] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2010, pp. 94–101.

[9] A. Vinciarelli and M. Pantic, "Techware: www.sspnet.eu: A Web Portal for Social Signal Processing [Best of the Web]," *IEEE Signal Processing Magazine*, Vol. 27, No. 4, pp. 142–144, 2010.

[10] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, "AVEC 2013 - the continuous audio/visual emotion and depression recognition challange," in *Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2013, pp. 1–8.

[11] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, "AVEC 2014: 3D Dimensional Affect and Depression Recognition Challenge," in *ACM International Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2014, pp. 3–10.

[12] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. T. Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "AVEC 2016 – Depression, Mood, and Emotion Recognition Workshop and Challenge," in *ACM International Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2016, pp. 3–10.

[13] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, and M. Pantic, "AVEC 2017 – Real-life Depression, and Affect Recognition Workshop and Challenge," in *ACM International Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2017, pp. 1–7.

[14] FP7-ICT SSPNet, "Social Signal Processing Network," 2018. [Online]. Available: https://cordis.europa.eu/project/rcn/89261{_}en.html

[15] R. W. Picard, "Affective Computing for HCI," in *ACM International on Human-Computer Interaction: Ergonomics and User Interfaces*, 1999, pp. 829–833.

[16] S. Narayanan and P. G. Georgiou, "Behavioral Signal Processing: Deriving Human Behavioral Informatics From Speech and Language," *Proceedings of the IEEE*, Vol. 101, No. 5, pp. 1203–1233, 2013.

[17] M. Valstar, "Automatic Behaviour Understanding in Medicine," in *Workshop on Roadmapping the Future of Multimodal Interaction Research Including Business Opportunities and Challenges*, 2014, pp. 57–60.

[18] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5)*, 5th ed. Arlington, VA: American Psychiatric Publishing, 2013.

[19] K. E. Stephan and C. Mathys, "Computational approaches to psychiatry," *Current Opinion in Neurobiology*, Vol. 25, pp. 85–92, 2014.

[20] M. Hamilton, "A rating scale for Depression," *Journal of Neurology, Neurosurgery, and Psychiatry*, Vol. 23, pp. 56–62, 1960.

[21] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. W. Williams, J. T. Berry, and A. H. Mokdad, "The PHQ-8 as a measure of current depression in the general population," *Journal of Affective Disorders*, Vol. 114, No. 1–3, pp. 163–173, 2009.

[22] R. C. Young, J. T. Biggs, V. E. Ziegler, and D. A. Meyer, "A rating scale for mania: reliability, validity, and sensitivity," *The British Journal of Psychiatry: The Journal of Mental Science*, Vol. 133, pp. 429–435, 1978.

[23] Psychology Tools, "Young Mania Rating Scale (online form)," 2018. [Online]. Available: https://psychology-tools.com/young-mania-rating-scale

[24] Z. S. Syed, J. Schroeter, K. Sidorov, and D. Marshall, "Computational Paralinguistics: Automatic Assessment of Emotions, Mood, and Behavioural State from Acoustics of Speech," in *INTERSPEECH*, 2018, pp. 511–515.

[25] L. R. Snowden, "Bias in mental health assessment and intervention: Theory and evidence," *American Journal of Public Health*, Vol. 93, No. 2, pp. 239–243, 2003.

[26] G. Bedi, F. Carrillo, G. A. Cecchi, D. F. Slezak, M. Sigman, N. B. Mota, S. Ribeiro, D. C. Javitt, M. Copelli, and C. M. Corcoran, "Automated analysis of free speech predicts psychosis onset in high-risk youths," *Nature Partner Journal: Schizophrenia*, Vol. 1, pp. 1–7, 2015.

[27] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. De La Torre, "Detecting depression from facial actions and vocal prosody," in *International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2009, pp. 1–7.

[28] C. Solomon, M. F. Valstar, R. K. Morriss, and J. Crowe, "Objective Methods for Reliable Detection of Concealed Depression," *Frontiers in ICT*, Vol. 2, pp. 1–16, 2015.

[29] B. Schuller, F. Weninger, Y. Zhang, F. Ringeval, A. Batliner, S. Steidl, F. Eyben, E. Marchi, A. Vinciarelli, K. Scherer, M. Chetouani, and M. Mortillaro, "Affective and behavioural computing: Lessons learnt from the First Computational Paralinguistics Challenge," *Computer Speech and Language*, Vol. 1, No. 1, pp. 1–25, 2018.

[30] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," in *ACM international conference on Multimedia*, 2013, pp. 835–838.

[31] T. Drugman and Y. Stylianou, "Maximum voiced frequency estimation: Exploiting amplitude and phase spectra," *IEEE Signal Processing Letters*, Vol. 21, No. 10, pp. 1230–1234, 2014.

[32] T. Baltrusaitis, P. Robinson, and L. P. Morency, "OpenFace: An open source facial behavior analysis toolkit," in *IEEE Winter Conference on Applications of Computer Vision*, 2016, pp. 1–10.

[33] E. Ciftci, H. Kaya, H. Gulec, and A. A. Salah, "The Turkish Audio-Visual Bipolar Disorder Corpus," in *Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*, 2018, pp. 1–6.

[34] Z. S. Syed, K. Sidorov, and D. Marshall, "Depression Severity Prediction Based on Biomarkers of Psychomotor Retardation," in *ACM International Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2017, pp. 37–43.

[35] J. Gareth, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*, 1st ed. Springer-Verlag New York, 2013.

[36] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2007, pp. 1–8.

[37] Z. S. Shah, K. Sidorov, and D. Marshall, "Psychomotor Cues for Depression Screening," in *IEEE International Conference on Digital Signal Processing (DSP)*, 2017, pp. 1–5.

[38] A. Beck, R. Steer, and G. Brown, "Manual for the Beck Depression Inventory-II," *San Antonio, TX: Psychological Corporation*, pp. 1–82, 1996.

[39] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, H. Kaya, M. Schmitt, S. Amiriparian, D. Lalanne, N. Cummins, A. Michaud, E. Ciftci, H. Gulec, A. A. Salah, and M. Pantic, "AVEC 2018 Workshop and Challenge: Bipolar Disorder and Cross-Cultural Affect Recognition," in *Audio/Visual Emotion Challenge Workshop*, 2018, pp. 1–11.

[40] Imotions.com, "Facial Action Coding System (FACS) — A Visual Guidebook," 2018. [Online]. Available: https://imotions.com/blog/facial-action-coding-system/

[41] Max Planck Institute for Psycholinguistics, "ELAN," 2018. [Online]. Available: http://tla.mpi.nl/tools/tla-tools/elan/

[42] H. Brugman, A. Russel, and X. Nijmegen, "Annotating multi-media / multimodal resources with ELAN," in *In proceedings of LREC*, 2004, pp. 2065–2068.

[43] H. B. Mann and D. R. Whitney, "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other," *The Annals of Mathematical Statistics*, Vol. 18, No. 1, pp. 50–60, 1947.

[44] G. M. Sullivan and R. Feinn, "Using Effect Size - or Why the P Value Is Not Enough," *Journal of Graduate Medical Education*, Vol. 4, No. 3, pp. 279–82, 2012.

[45] C. O. Fritz, P. E. Morris, and J. J. Richler, "Effect size estimates: Current use, calculations, and interpretation," *Journal of Experimental Psychology: General*, Vol. 141, No. 1, pp. 2–18, 2012.

[46] R. Fusaroli, A. Lambrechts, D. Bang, D. M. Bowler, and S. B. Gaigg, "Is voice a marker for Autism spectrum disorder? A systematic review and meta-analysis," *Autism Research*, pp. 1–50, 2016.

[47] T. F. Yap, J. Epps, E. Ambikairajah, and E. H. C. Choi, "Voice source under cognitive load: Effects and classification," *Speech Communication*, Vol. 72, pp. 74–95, 2015.

[48] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, Vol. 44, No. 3, pp. 572–587, 2011.

[49] R. Kliper, S. Portuguese, and D. Weinshall, "Prosodic analysis of speech and the underlying mental state," in *Communications in Computer and Information Science*, Vol. 604, 2016, pp. 52–62.

[50] T. Bocklet, E. Noth, G. Stemmer, H. Ruzickova, and J. Rusz, "Detection of persons with Parkinson's disease by acoustic, vocal, and prosodic analysis," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011, pp. 478–483.

[51] J. Weiner, C. Herff, and T. Schultz, "Speech-based detection of Alzheimer's disease in conversational German," in *INTERSPEECH*, 2016, pp. 1938–1942.

[52] K. C. Fraser, F. Rudzicz, and G. Hirst, "Detecting late-life depression in Alzheimer's disease through analysis of speech and language," in *Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2016, pp. 1–11.

[53] D. K. Oller, P. Niyogi, S. Gray, J. a. Richards, J. Gilkerson, D. Xu, U. Yapanel, and S. F. Warren, "Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development," *National Academy of Sciences of the United States of America*, Vol. 107, No. 30, pp. 13 354–13 359, 2010.

[54] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete Time Signal Processing*, 3rd ed. Pearson, 2009.

[55] J. Proakis and D. Manolakis, *Digital Signal Processing*, 4th ed. Pearson, 2006.

[56] D. Mitrovic, M. Zeppelzauer, and C. Breiteneder, "Features for Content-Based Audio Retrieval," *Advances in Computers*, Vol. 78, No. 10, pp. 71–150, 2010.

[57] F. Alias, J. Socoro, and X. Sevillano, "A Review of Physical and Perceptual Feature Extraction Techniques for Speech, Music and Environmental Sounds," *Applied Sciences*, Vol. 6, No. 5, pp. 1–44, 2016.

[58] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andre, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Transactions on Affective Computing*, Vol. 7, No. 2, pp. 190–202, 2016.

[59] J.-T. Huang, J. Li, and Y. Gong, "An analysis of convolutional neural networks for speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4989–4993.

[60] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," *arXiv:1609.03499*, pp. 1–15, 2016.

[61] C. Candan, M. Alper Kutay, and H. M. Ozaktas, "The discrete fractional fourier transform," *IEEE Transactions on Signal Processing*, Vol. 48, No. 9, pp. 1329–1337, 2000.

[62] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP — A collaborative voice analysis repository for speech technologies," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 960–964.

[63] Y.-L. Shue, P. Keating, and C. Vicenik, "VOICESAUCE: A program for voice analysis," *The Journal of the Acoustical Society of America*, Vol. 126, No. 8, p. 2221, 2009.

[64] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *INTERSPEECH*, 2013, pp. 148–152.

[65] T. Giannakopoulos and A. Pikrakis, *Introduction to Audio Analysis*, 1st ed. Academic Press Inc. (London) Ltd., 2014.

[66] M. A. Sayette, J. F. Cohn, J. M. Wertz, M. A. Perrott, and D. J. Parrott, "A psychometric evaluation of the facial action coding system for assessing spontaneous expression," *Journal of Nonverbal Behavior*, Vol. 25, No. 3, pp. 167–185, 2001.

[67] J. F. Cohn and F. D. L. Torre, "Automated Face Analysis for Affective Computing," in *Handbook of Affective Computing*. Oxford University Press, 2014.

[68] T. Guha, Z. Yang, A. Ramakrishna, R. B. Grossman, D. Hedley, S. Lee, and S. S. Narayanan, "On quantifying facial expression-related atypicality of children with Autism Spectrum Disorder," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 803–807.

[69] M. D. Coco, M. Leo, P. Carcagni, P. Spagnolo, P. L. Mazzeo, M. Bernava, F. Marino, G. Pioggia, and C. Distante, "A Computer Vision based Approach for Understanding Emotional Involvements in Children with Autism Spectrum Disorders," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1401–1407.

[70] M. D. Samad, N. Diawara, J. L. Bobzien, J. W. Harrington, M. A. Witherow, and K. M. Iftekharuddin, "A Feasibility Study of Autism Behavioral Markers in Spontaneous Facial, Visual, and Hand Movement Response Data," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, Vol. 26, No. 2, pp. 353–361, 2018.

[71] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. Mavadati, and D. P. Rosenwald, "Social risk and depression: Evidence from manual and automatic facial expression analysis," in *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, 2013, pp. 1–8.

[72] T. Tron, A. Peled, A. Grinsphoon, and D. Weinshall, "Automated Facial Expressions Analysis in Schizophrenia: A Continuous Dynamic Approach," in *International Conference on Pervasive Computing Paradigms for Mental Health (MindCare)*, 2016, pp. 72–81.

[73] ——, "Facial expressions and flat affect in schizophrenia, automatic analysis from depth camera data," in *IEEE International Conference on Biomedical and Health Informatics (EMBS)*, 2016, pp. 220–223.

[74] S. Vijay, T. Baltrusaitis, L. Pennant, D. Ongur, J. T. Baker, and L.-P. Morency, "Computational study of psychosis symptoms and facial expressions," in *ACM Conference on Human Factors in Computing Systems (CHI)*, 2016, pp. 1–4.

[75] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 37, No. 6, pp. 1113–1133, 2014.

[76] P. Viola and M. Jones, "Robust Real-Time Face Detection," *International Journal of Computer Vision*, Vol. 57, pp. 137–154, 2004.

[77] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on Local Binary Patterns: A comprehensive study," *Image and Vision Computing*, Vol. 27, No. 6, pp. 803–816, 2009.

[78] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2005, pp. 886–893.

[79] P. Carcagni, M. Del Coco, M. Leo, and C. Distante, "Facial expression recognition and histograms of oriented gradients: A comprehensive study," *SpringerPlus*, Vol. 4, No. 1, pp. 1–25, 2015.

[80] A. Dhall, A. Asthana, R. Goecke, and T. Gedeon, "Emotion recognition using PHOG and LPQ features," in *IEEE International Conference on Automatic Face and Gesture Recognition and Workshops*, 2011, pp. 878–883.

[81] G. Zhao and M. Pietikainen, "Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 29, No. 6, pp. 915–928, 2007.

[82] D. E. King, "Dlib-ml: A Machine Learning Toolkit," *Journal of Machine Learning Research*, Vol. 10, pp. 1755–1758, 2009.

[83] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Incremental face alignment in the wild," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1859–1866.

[84] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett, "The computer expression recognition toolbox (CERT)," in *IEEE International Conference on Automatic Face and Gesture Recognition and Workshops*, 2011, pp. 298–305.

[85] F. De la Torre, Wen-Sheng Chu, Xuehan Xiong, F. Vicente, Xiaoyu Ding, and J. Cohn, "IntraFace," in *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 2015, pp. 1–8.

[86] B. Martinez, M. F. Valstar, B. Jiang, and M. Pantic, "Automatic Analysis of Facial Actions: A Survey," *IEEE Transactions on Affective Computing*, Vol. 1, No. 1, pp. 1–27, 2017.

[87] A. Krizhevsky, I. Sutskever, and H. Geoffrey E., "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2012, pp. 1–9.

[88] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *International Conference on Learning Representations (ICRL)*, pp. 1–14, 2015.

[89] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9.

[90] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[91] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 815–823.

[92] L. Yang, H. Sahli, X. Xia, E. Pei, M. C. Oveneke, and D. Jiang, "Hybrid Depression Classification and Estimation from Audio Video and Text Information," in *ACM International Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2017, pp. 1–7.

[93] T. Althoff, K. Clark, and J. Leskovec, "Large-scale Analysis of Counseling Conversations: An Application of Natural Language Processing to Mental Health," *Transactions of the Association for Computational Linguistics*, Vol. 4, pp. 463–476, 2016.

[94] R. A. Calvo, D. N. Milne, M. S. Hussain, and H. Christensen, "Natural language processing in mental health applications using non-clinical texts," *Natural Language Engineering*, Vol. 23, No. 5, pp. 649–685, 2017.

[95] D. A. Reynolds and R. C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models," *IEEE Transactions on Speech and Audio Processing*, Vol. 3, No. 1, pp. 72–83, 1995.

[96] D. C. Smith and K. A. Kornelson, "A Comparison of Fisher Vectors and Gaussian Supervectors for Document Versus Non-document Image Classification," in *SPIE 8856, Applications of Digital Image Processing XXXVI*, Vol. 8856, 2013, pp. 1–12.

[97] G. Csurka, C. Dance, L. Fan, J. Willamowski, and B. Cedric, "Visual categorization with bag of keypoints," in *International Workshop on Statistical Learning in Computer Vision*, 2004, pp. 1–22.

[98] J. Sanchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: Theory and practice," *International Journal of Computer Vision*, Vol. 105, No. 3, pp. 222–245, 2013.

[99] F. Perronnin, J. Sanchez, and T. Mensink, "Improving the Fisher kernel for large-scale image classification," in *Lecture Notes in Computer Science*, Vol. 6314, 2010, pp. 143–156.

[100] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, Vol. 3, pp. 1157–1182, 2003.

[101] L. van der Maaten, E. Postma, and J. van den Herik, "Dimensionality Reduction: A Comparative Review," *Journal of Machine Learning Research*, Vol. 10, No. February, pp. 1–41, 2009.

[102] H. Kaya, F. Eyben, A. A. Salah, and B. Schuller, "CCA based feature selection with application to continuous depression recognition from acoustic speech features," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 3729–3733.

[103] R. Rosipal and N. Kr, "Overview and Recent Advances in Partial Least Squares," *Subspace, Latent Structure and Feature Selection*, Vol. 3940, pp. 34–51, 2006.

[104] S. Alghowinem, R. Goecke, M. Wagner, G. Parkerx, and M. Breakspear, "Head Pose and Movement Analysis as an Indicator of Depression," in *International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2013, pp. 283–288.

[105] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers and Electrical Engineering*, Vol. 40, No. 1, pp. 16–28, 2014.

[106] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 27, No. 8, pp. 1226–1238, 2005.

[107] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *International workshop on Machine Learning*, 1992, pp. 249–256.

[108] J. Kittler, "A Framework for Classifier Fusion: Is It Still Needed?" in *Advances in Pattern Recognition*, F. J. Ferri, J. M. Inesta, A. Amin, and

P. Pudil, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000, pp. 45–56.

[109] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, T. Gedeon, M. Breakspear, and G. Parker, "A comparative study of different classifiers for detecting depression from spontaneous speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 8022–8026.

[110] A. Dhall and R. Goecke, "A temporally piece-wise fisher vector approach for depression analysis," in *International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2015, pp. 255–259.

[111] A. Pampouchidou, O. Simantiraki, C. M. Vazakopoulou, C. Chatzaki, M. Pediaditis, A. Maridaki, K. Marias, P. Simos, F. Yang, F. Meriaudeau, and M. Tsiknakis, "Facial geometry and speech analysis for depression detection," in *IEEE Engineering in Medicine and Biology Society (EMBC)*, 2017, pp. 1433–1436.

[112] H. Kaya, A. A. Karpov, and A. A. Salah, "Fisher Vectors with Cascaded Normalization for Paralinguistic Analysis," in *INTERSPEECH*, 2015, pp. 909–913.

[113] M. Kachele, M. Glodek, D. Zharkov, S. Meudt, and F. Schwenker, "Fusion of audio-visual features using hierarchical classifier systems for the recognition of affective states and the state of depression," in *ACM International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, 2014, pp. 671–678.

[114] R. E. Kalman, "A New Approach to Linear Filtering and Prediction Problems," *Journal of Basic Engineering*, Vol. 82, pp. 35–45, 1960.

[115] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Workshop on Computational Learning Theory (COLT)*, 1992, pp. 144–152.

[116] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, Vol. 20, No. 3, pp. 273–297, 1995.

[117] H. Drucker, C. J. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," *Advances in Neural Information Processing Systems (NIPS)*, Vol. 9, pp. 155–161, 1997.

[118] T. K. Ho, "Random Decision Forests," in *International Conference on Document Analysis and Recognition*, 1995, pp. 278–282.

[119] ——, "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 8, pp. 832–844, 1998.

[120] C. M. Bishop, *Pattern Recognition and Machine Learning*, 1st ed. Springer-Verlag New York, 2006.

[121] M. E. Tipping, "Sparse Bayesian Learning and the Relevance Vector Machine," *Journal of Machine Learning Research*, Vol. 1, pp. 211–244, 2001.

[122] G.-b. Huang, Q.-y. Zhu, and C.-k. Siew, "Extreme Learning Machine : A New Learning Scheme of Feedforward Neural Networks," *IEEE International Joint Conference on Neural Networks*, Vol. 2, pp. 985–990, 2004.

[123] G.-B. Huang, Q.-y. Zhu, and C.-k. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, Vol. 70, No. 1-3, pp. 489–501, 2006.

[124] D. R. Cox, "The regression analysis of binary sequences," *Journal of the Royal Statistical Society: Series B*, Vol. 20, No. 2, pp. 215–242, 1958.

[125] S. Scherer, G. Stratou, J. Gratch, and L. P. Morency, "Investigating voice quality as a speaker-independent indicator of depression and PTSD," in *INTERSPEECH*, 2013, pp. 847–851.

[126] S. Scherer, G. Stratou, and L.-P. Morency, "Audiovisual behavior descriptors for depression assessment," in *ACM International Conference on Multimodal Interaction (ICMI)*, 2013, pp. 135–140.

[127] F. Honig, A. Batliner, E. Noth, S. Schnieder, and J. Krajewski, "Automatic modelling of depressed speech: relevant features and relevance of gender," in *INTERSPEECH*, 2014, pp. 1248–1252.

[128] J. R. Williamson, T. F. Quatieri, B. S. Helfer, G. Ciccarelli, and D. D. Mehta, "Vocal and Facial Biomarkers of Depression Based on Motor Incoordination and Timing," in *ACM International Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2014, pp. 65–72.

[129] S. Alghowinem, R. Goecke, J. Epps, M. Wagner, and J. Cohn, "Cross-Cultural Depression Recognition from Vocal Biomarkers," in *INTERSPEECH*, 2016, pp. 1–5.

[130] J. R. Williamson, E. Godoy, M. Cha, A. Schwarzentruber, P. Khorrami, Y. Gwon, H.-T. Kung, C. Dagli, and T. F. Quatieri, "Detecting Depression Using Vocal, Facial and Semantic Communication Cues," in *ACM*

International Workshop on Audio/Visual Emotion Challenge (AVEC), 2016, pp. 11–18.

[131] Y. Gong and C. Poellabauer, "Topic Modeling Based Multi-modal Depression Detection," in *ACM International Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2017, pp. 1–7.

[132] G. He, J. Chen, X. Liu, and M. Li, "The SYSU system for CCPR 2016 multimodal emotion recognition challenge," in *Communications in Computer and Information Science*, 2016, pp. 707–720.

[133] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, and G. Anbarjafari, "Audio-Visual Emotion Recognition in Video Clips," *IEEE Transactions on Affective Computing*, Vol. 1, No. 1, pp. 1–17, 2017.

[134] B. S. Simone Hantke, Hesam Sagha, Nicholas Cummins, "Emotional Speech of Mentally and Physically Disabled Individuals: Introducing the EmotAsS Database and First Findings," in *INTERSPEECH*, 2017, pp. 3137–3141.

[135] B. Nojavanasghari, T. Baltrušaitis, C. E. Hughes, and L.-P. Morency, "EmoReact: a multimodal approach and dataset for recognizing emotional responses in children," in *ACM International Conference on Multimodal Interaction (ICMI)*, 2016, pp. 137–144.

[136] M. Brilman and S. Scherer, "A Multimodal Predictive Model of Successful Debaters or How I Learned to Sway Votes," in *ACM International Conference on Multimedia (MM)*, 2015, pp. 149–158.

[137] T. Wortwein, M. Chollet, B. Schauerte, L.-P. Morency, R. Stiefelhagen, and S. Scherer, "Multimodal Public Speaking Performance Assessment," in *ACM on International Conference on Multimodal Interaction (ICMI)*, 2015, pp. 43–50.

[138] I. Naim, M. I. Tanveer, D. Gildea, and M. E. Hoque, "Automated prediction and analysis of job interview performance: The role of what you say and how you say it," in *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 2015, pp. 1–6.

[139] I. Naim, M. I. Tanveer, D. Gildea, and E. Hoque, "Automated Analysis and Prediction of Job Interview Performance," *IEEE Transactions on Affective Computing*, Vol. 9, No. 2, pp. 191–204, 2016.

[140] J. F. Santos, N. Brosh, T. H. Falk, L. Zwaigenbaum, S. E. Bryson, W. Roberts, I. M. Smith, P. Szatmari, and J. A. Brian, "Very early detection of Autism Spectrum Disorders based on acoustic analysis of

pre-verbal vocalizations of 18-month old toddlers," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 7567–7571.

[141] M. Asgari, A. Bayestehtashk, and I. Shafran, "Robust and accurate features for detecting and diagnosing autism spectrum disorders." in *INTERSPEECH*, 2013, pp. 191–194.

[142] E. Marchi, B. Schuller, S. Baron-Cohen, O. Golan, S. Bolte, P. Arora, and R. Hab-Umbach, "Typicality and emotion in the voice of children with autism spectrum condition: Evidence across three languages," in *INTERSPEECH*, 2015, pp. 115–119.

[143] M. Schmitt, E. Marchi, F. Ringeval, and B. Schuller, "Towards Cross-lingual Automatic Diagnosis of Autism Spectrum Condition in Children's Voices," in *ITG-Fachbericht 267: Speech Communication*, 2016, pp. 264–268.

[144] A. Baird, S. Amiriparian, N. Cummins, A. M. Alcorn, A. Batliner, S. Pugachevskiy, M. Freitag, M. Gerczuk, and B. Schuller, "Automatic Classification of Autistic Child Vocalisations: A Novel Database and Results," in *INTERSPEECH*, 2017, pp. 1–5.

[145] F. B. Pokorny, B. W. Schuller, P. B. Marschik, R. Brueckner, P. Nystrom, N. Cummins, S. Bolte, C. Einspieler, and T. Falck-Ytter, "Earlier Identification of Children with Autism Spectrum Disorder: An Automatic Vocalisation-based Approach," in *INTERSPEECH*, 2017, pp. 309–313.

[146] B. Schuller, S. Steidl, A. Batliner, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom, A. S. Warlaumont, G. Hidalgo, S. Schnieder, C. Heiser, W. Hohenhorst, M. Herzog, M. Schmitt, K. Qian, Y. Zhang, G. Trigeorgis, P. Tzirakis, and S. Zafeiriou, "The INTERSPEECH 2017 Computational Paralinguistics Challenge: Addressee, Cold and Snoring," in *INTERSPEECH*, 2017, pp. 1–5.

[147] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, A. Baird, and B. Schuller, "Snore sound classification using image-based deep spectrum features," in *INTERSPEECH*, 2017, pp. 3512–3516.

[148] B. W. Schuller, S. Steidl, A. Batliner, P. B. Marschik, H. Baumeister, F. Dong, S. Hantke, F. Pokorny, E.-M. Rathner, K. D. Bartl-Pokorny, C. Einspieler, D. Zhang, A. Baird, S. Amiriparian, K. Qian, Z. Ren, M. Schmitt, P. Tzirakis, and S. Zafeiriou, "The INTERSPEECH 2018 Computational Paralinguistics Challenge: Atypical and Self-Assessed Affect, Crying and Heart Beats," in *INTERSPEECH*, 2018, pp. 1–5.

[149] L. Yang, D. Jiang, L. He, E. Pei, M. C. Oveneke, and H. Sahli, "Decision Tree Based Depression Classification from Audio Video and Language Information," in *ACM International Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2016, pp. 89–96.

[150] A. Pampouchidou, O. Simantiraki, A. Fazlollahi, M. Pediaditis, D. Manousos, A. Roniotis, G. Giannakakis, F. Meriaudeau, P. Simos, K. Marias, F. Yang, and M. Tsiknakis, "Depression Assessment by Fusing High and Low Level Features from Audio, Video, and Text," in *ACM International Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2016, pp. 27–34.

[151] B. Stasak, J. Epps, and R. Goecke, "Elicitation design for acoustic depression classification: An investigation of articulation effort, linguistic complexity, and word affect," in *INTERSPEECH*, 2017, pp. 834–838.

[152] M. Nasir, A. Jati, P. G. Shivakumar, S. Nallan Chakravarthula, and P. Georgiou, "Multimodal and Multiresolution Depression Detection from Speech and Facial Landmark Features," in *ACM International Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2016, pp. 43–50.

[153] B. Sun, Y. Zhang, J. He, L. Yu, Q. Xu, D. Li, and Z. Wang, "A Random Forest Regression Method With Selected-Text Feature For Depression Assessment," in *ACM International Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2017, pp. 1–7.

[154] M. Senoussaoui, M. Sarria-Paja, J. F. Santos, and T. H. Falk, "Model Fusion for Multimodal Depression Classification and Level Detection," in *ACM International Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2014, pp. 57–63.

[155] Z. Huang, B. Stasak, T. Dang, K. Wataraka Gamage, P. Le, V. Sethu, and J. Epps, "Staircase Regression in OA RVM, Data Selection and Gender Dependency in AVEC 2016," in *ACM International Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2016, pp. 19–26.

[156] B. Stasak, J. Epps, N. Cummins, and R. Goecke, "An investigation of emotional speech in depression classification," in *INTERSPEECH*, 2016, pp. 485–489.

[157] G. Kiss and K. Vicsi, "Mono- and multi-lingual depression prediction based on speech processing," *International Journal of Speech Technology*, Vol. 20, No. 4, pp. 919–935, 2017.

[158] T. Dang, B. Stasak, Z. Huang, S. Jayawardena, M. Atcheson, M. Hayat, P. Le, V. Sethu, R. Goecke, and J. Epps, "Investigating Word Affect

Features and Fusion of Probabilistic Predictions Incorporating Uncertainty in AVEC 2017," in *ACM International Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2017, pp. 1–7.

[159] J. Egede, M. Valstar, and B. Martinez, "Fusing Deep Learned and Hand-Crafted Features of Appearance, Shape, and Dynamics for Automatic Pain Estimation," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2017, pp. 1–8.

[160] H. Kaya, A. Karpov, and A. Salah, "Robust acoustic emotion recognition based on cascaded normalization and extreme learning machines," in *Advances in Neural Networks*, 2016, pp. 115–123.

[161] H. Kaya and A. A. Karpov, "Introducing weighted kernel classifiers for handling imbalanced paralinguistic corpora: Snoring, addressee and cold," in *INTERSPEECH*, 2017, pp. 3527–3531.

[162] N. Cummins, J. Epps, M. Breakspear, and R. Goecke, "An investigation of depressed speech detection: Features and normalization," in *INTERSPEECH*, 2011, pp. 2997–3000.

[163] Y. Yang, C. Fairbairn, and J. F. Cohn, "Detecting depression severity from vocal prosody," *IEEE Transactions on Affective Computing*, Vol. 4, No. 2, pp. 142–150, 2013.

[164] V. Mitra, E. Shriberg, M. Mclaren, A. Kathol, C. Richey, D. Vergyri, and M. Graciarena, "The SRI AVEC-2014 Evaluation System," in *ACM International Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2014, pp. 93–101.

[165] S. Chen, Y. Dian, X. Li, X. Lin, Q. Jin, H. Liu, and L. Lu, "Emotion Recognition in Videos via Fusing Multimodal Features," in *Chinese Conference on Pattern Recognition (CCPR)*, 2016, pp. 632–644.

[166] H. Dibeklioglu, Z. Hammal, and J. F. Cohn, "Dynamic Multimodal Measurement of Depression Severity Using Deep Autoencoding," *IEEE Journal of Biomedical and Health Informatics*, Vol. 22, No. 2, pp. 525–536, 2018.

[167] C. P. Chen, X. H. Tseng, S. S. F. Gau, and C. C. Lee, "Computing multimodal dyadic behaviors during spontaneous diagnosis interviews toward automatic categorization of autism spectrum disorder," in *INTERSPEECH*, 2017, pp. 2361–2365.

[168] O. Celiktutan and H. Gunes, "Automatic Prediction of Impressions in Time and across Varying Context: Personality, Attractiveness and Likeability," *IEEE Transactions on Affective Computing*, Vol. 1, No. 1, pp. 29–42, 2017.

[169] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A Practical Guide to Support Vector Classification," *BJU international*, Vol. 101, No. 1, pp. 1396–400, 2008.

[170] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD Explorations Newsletter*, Vol. 11, No. 1, pp. 10–18, 2009.

[171] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A Library for Large Linear Classification," *Journal of Machine Learning Research*, Vol. 9, pp. 1871–1874, 2008.

[172] S. Wold, M. Sjostrom, and L. Eriksson, "PLS-regression: A basic tool of chemometrics," *Chemometrics and Intelligent Laboratory Systems*, Vol. 58, No. 2, pp. 109–130, 2001.

[173] S. Scherer, G. Stratou, M. Mahmoud, J. Boberg, J. Gratch, A. Rizzo, and L. P. Morency, "Automatic behavior descriptors for psychological disorder analysis," in *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, 2013, pp. 1–8.

[174] S. Alghowinem, R. Goecke, M. Wagner, G. Parker, and M. Breakspear, "Eye movement analysis for depression detection," in *IEEE International Conference on Image Processing*, 2013, pp. 4220–4224.

[175] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Hyett, G. Parker, and M. Breakspear, "Multimodal Depression Detection: Fusion Analysis of Paralinguistic, Head Pose and Eye Gaze Behaviors," *IEEE Transactions on Affective Computing*, pp. 1–14, 2016.

[176] J. Gideon, E. M. Provost, and M. McInnis, "Mood state prediction from speech of varying acoustic quality for individuals with bipolar disorder," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 2359–2363.

[177] G. M. Lucas, J. Gratch, S. Scherer, J. Boberg, and G. Stratou, "Towards an Affective Interface for Assessment of Psychological Distress," in *International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2015, pp. 539–545.

[178] D. DeVault, K. Georgila, R. Artstein, F. Morbini, D. Traum, S. Scherer, A. Rizzo, and L.-P. Morency, "Verbal indicators of psychological distress in interactive dialogue with a virtual human," in *Special Interest Group on Discourse and Dialogue (SIGDIAL) Conference*, 2013, pp. 193–202.

[179] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. M. Mavadati, Z. Hammal, and D. P. Rosenwald, "Nonverbal social withdrawal in depression: Evidence

from manual and automatic analyses," *Image and Vision Computing*, Vol. 32, No. 10, pp. 641–647, 2014.

[180] E. Laksana, T. Baltrusaitis, L.-P. Morency, and J. P. Pestian, "Investigating Facial Behavior Indicators of Suicidal Ideation," in *International Conference on Automatic Face and Gesture Recognition*, 2017, pp. 1–8.

[181] R. Paul, A. Augustyn, A. Klin, and F. R. Volkmar, "Perception and production of prosody by speakers with autism spectrum disorders," *Journal of Autism and Developmental Disorders*, Vol. 35, No. 2, pp. 205–220, 2005.

[182] A.-M. R. DePape, A. Chen, G. B. C. Hall, and L. J. Trainor, "Use of Prosody and Information Structure in High Functioning Adults with Autism in Relation to Language Ability," *Frontiers in Psychology*, Vol. 3, pp. 1–13, 2012.

[183] A. Guidi, J. Schoentgen, G. Bertschy, C. Gentili, E. P. Scilingo, and N. Vanello, "Features of vocal frequency contour and speech rhythm in bipolar disorder," *Biomedical Signal Processing and Control*, Vol. 37, pp. 23–31, 2017.

[184] J. Zhang, Z. Pan, C. Gui, T. Xue, Y. Lin, J. Zhu, and D. Cui, "Analysis on speech signal features of manic patients," *Journal of Psychiatric Research*, Vol. 98, pp. 59–63, 2018.

[185] B. Yu, T. F. Quatieri, J. R. Williamson, and J. C. Mundt, "Cognitive impairment prediction in the elderly based on vocal biomarkers," in *INTERSPEECH*, 2015, pp. 3734–3738.

[186] G. Stratou, S. Scherer, J. Gratch, and L. P. Morency, "Automatic Nonverbal Behavior Indicators of Depression and PTSD: Exploring Gender Differences," in *IEEE Affective Computing and Intelligent Interaction (ACII)*, 2013, pp. 147–152.

[187] ——, "Automatic nonverbal behavior indicators of depression and PTSD: the effect of gender," *Journal on Multimodal User Interfaces*, Vol. 9, No. 1, pp. 17–29, 2015.

[188] B. Stasak, J. Epps, and N. Cummins, "Depression Prediction Via Acoustic Analysis of Formulaic Word Fillers," in *Australasian International Conference on Speech Science and Technology*, 2016, pp. 277–280.

[189] S. I. Levitan, G. An, M. Ma, R. Levitan, A. Rosenberg, and J. Hirschberg, "Combining acoustic-prosodic, lexical, and phonotactic features for automatic deception detection," in *INTERSPEECH*, 2016, pp. 1–5.

[190] T. Wortwein, L.-P. Morency, and S. Scherer, "Automatic Assessment and Analysis of Public Speaking Anxiety: A Virtual Audience Case Study," in *IEEE Affective Computing and Intelligent Interaction (ACII)*, 2015, pp. 1–7.

[191] M. Chollet, L.-p. Morency, A. Shapiro, S. Scherer, and L. Angeles, "Exploring Feedback Strategies to Improve Public Speaking: An Interactive Virtual Audience Framework," in *ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, 2015, pp. 1143–1154.

[192] S. Scherer, G. Stratou, G. Lucas, M. Mahmoud, J. Boberg, J. Gratch, A. Rizzo, and L. P. Morency, "Automatic audiovisual behavior descriptors for psychological disorder analysis," *Image and Vision Computing*, Vol. 32, No. 10, pp. 648–658, 2014.

[193] V. Mitra, E. Shriberg, D. Vergyri, B. Knoth, and R. M. Salomon, "Cross-corpus depression prediction from speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4769–4773.

[194] D. Bone, M. Black, and C. Lee, "Spontaneous-Speech Acoustic-Prosodic Features of Children with Autism and the Interacting Psychologist," *INTERSPEECH*, pp. 3–6, 2012.

[195] D. Bone, C.-C. Lee, M. P. Black, M. E. Williams, S. Lee, P. Levitt, and S. Narayanan, "The psychologist as an interlocutor in autism spectrum disorder assessment: insights from a study of spontaneous prosody," *Journal of Speech, Language, and Hearing Research*, Vol. 57, No. 4, pp. 1162–1177, 2014.

[196] T. Wortwein, T. Baltrusaitis, E. Laksana, L. Pennant, E. S. Liebson, D. Ongur, J. T. Baker, and L.-P. Morency, "Computational Analysis of Acoustic Descriptors in Psychotic Patients," in *INTERSPEECH*, 2017, pp. 3256–3260.

[197] M. K. Wolters, L. Ferrini, E. Farrow, A. S. Tatar, and C. D. Burton, "Tracking Depressed Mood Using Speech Pause Patterns," *International Congress of Phonetic Sciences*, pp. 1–5, 2015.

[198] R. L. Horwitz-Martin, T. F. Quatieri, E. Godoy, and J. R. Williamson, "A vocal modulation model with application to predicting depression severity," in *IEEE International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, 2016, pp. 1–7.

[199] A. S. Cohen, J. E. McGovern, T. J. Dinzeo, and M. A. Covington, "Speech Deficits in Serious mental Illness: A Cognitive Resource Issue?" *Schizophrenia research*, Vol. 160, No. 0, pp. 173–179, 2014.

[200] M. Asgari and I. Shafran, "Improvements to harmonic model for extracting better speech features in clinical applications," *Computer Speech & Language*, Vol. 47, pp. 298–313, 2017.

[201] S. Hantke, C. Cohrs, M. Schmitt, B. Tannert, F. Lutkebohmert, M. Detmers, H. Schelhowe, and B. Schuller, "Introducing an Emotion-Driven Assistance System for Cognitively Impaired Individuals," in *International Conference on Computers Helping People with Special Needs*, 2018, pp. 486–494.

[202] C. Sobin and H. A. Sackeim, "Psychomotor symptoms of depression," *American Journal of Psychiatry*, Vol. 154, No. 1, pp. 4–17, 1997.

[203] S. Alghowinem, "From joyous to clinically depressed: Mood detection using multimodal analysis of a person's appearance and speech," in *International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2013, pp. 648–653.

[204] S. Alghowinem, R. Goecke, J. F. Cohn, M. Wagner, G. Parker, and M. Breakspear, "Cross-cultural detection of depression from nonverbal behaviour," in *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, 2015, pp. 1–8.

[205] D. F. DeMenthon and L. S. Davis, "Model-based object pose in 25 lines of code," *International Journal of Computer Vision*, Vol. 15, No. 1-2, pp. 123–141, 1995.

[206] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 23, No. 6, pp. 681–685, 2001.

[207] H. Dibeklioğlu, Z. Hammal, Y. Yang, and J. F. Cohn, "Multimodal Detection of Depression in Clinical Interviews," in *ACM International Conference on Multimodal Interaction (ICMI)*, 2015, pp. 307–310.

[208] J. W. Davis, "Hierarchical motion history images for recognizing human motion," in *IEEE Workshop on Detection and Recognition of Events in Video*, 2001, pp. 1–8.

[209] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, R. Szeliski, S. S. Beauchemin, J. L. Barron, J. Campbell, R. Sukthankar, I. Nourbakhsh, S. Dickinson, M. Pelillo, R. Zabih, B. Galvin, B. Mccane, K. Novins, D. Mason, S. Mills, N. L. Johnson, T. W. Anderson, A. I. Laboratory, A. I. Laboratory, H. Liu, T.-h. Hong, M. Herman, T. Camus, B. Mccane, B. Galvin, K. Novins, D. Crannitch, B. Galvin, A. A. Shafie, F. Hafiz, M. H. Ali, D. V. We, and I. S. While, "Determining Optical Flow," *Artificial Intelligence*, Vol. 17, No. 3, pp. 185–203, 1998.

[210] J. R. Williamson, T. F. Quatieri, B. S. Helfer, R. Horwitz, B. Yu, and D. D. Mehta, "Vocal Biomarkers of Depression Based on Motor Incoordination," in *ACM International Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2013, pp. 41–48.

[211] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. Traum, A. S. Rizzo, and L.-P. Morency, "The Distress Analysis Interview Corpus of Human and Computer Interviews," in *International Conference on Language Resources and Evaluation (LREC)*, 2014, pp. 3123–3128.

[212] J. E. Overall and D. R. Gorham, "The Brief Psychiatric Rating Scale," *Psychological Reports*, Vol. 10, No. 3, pp. 799–812, 1962.

[213] S. R. Kay, A. Fiszbein, and L. A. Opler, "The positive and negative syndrome scale (PANSS) for schizophrenia," *Schizophrenia Bulletin*, Vol. 13, No. 2, pp. 261–276, 1987.

[214] Emotient, "Emotient FACET toolbox," 2018. [Online]. Available: https://imotions.com/facial-expressions/

[215] R. Horwitz, T. F. Quatieri, B. S. Helfer, B. Yu, J. R. Williamson, and J. Mundt, "On the relative importance of vocal source, system, and prosody in human depression," in *IEEE International Conference on Body Sensor Networks (BSN)*, 2013, pp. 1–6.

[216] M. S. De Bodt, M. E. Hernandez-Diaz Huici, and P. H. Van De Heyning, "Intelligibility as a linear combination of dimensions in dysarthric speech," *Journal of Communication Disorders*, Vol. 35, No. 3, pp. 283–292, 2002.

[217] K. Johnson, "Acoustic and Auditory Phonetics," *Phonetica*, Vol. 61, No. 1, pp. 56–58, 2004.

[218] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, Vol. 71, pp. 10–49, 2015.

[219] B. S. Helfer, T. F. Quatieri, J. R. Williamson, D. D. Mehta, R. Horwitz, and B. Yu, "Classification of depression state based on articulatory precision," in *INTERSPEECH*, 2013, pp. 2172–2176.

[220] A. Wingfield and P. A. Tun, "Cognitive Supports and Cognitive Constraints on Comprehension of Spoken Language," *Journal of the American Academy of Audiology*, Vol. 18, No. 7, pp. 548–558, 2007.

[221] E. Moore, M. Clements, J. Peifer, and L. Weisser, "Analysis of prosodic variation in speech for clinical depression," in *IEEE Engineering in Medicine and Biology Society (EMBC)*, Vol. 3, 2003, pp. 2925–2928.

[222] J. M. Girard and J. F. Cohn, "Automated Audiovisual Depression Analysis," *Current Opinion in Psychology*, Vol. 4, pp. 75–79, 2014.

[223] R. Gupta, N. Malandrakis, B. Xiao, T. Guha, M. Van Segbroeck, M. Black, A. Potamianos, and S. Narayanan, "Multimodal Prediction of Affective Dimensions and Depression in Human-Computer Interactions," in *ACM International Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2014, pp. 33–40.

[224] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Breakspear, and G. Parker, "Detecting depression: A comparison between spontaneous and read speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 7547–7551.

[225] A. Y. Hussenbocus, M. Lech, and N. B. Allen, "Statistical differences in speech acoustics of major depressed and non-depressed adolescents," in *International Conference on Signal Processing and Communication Systems*, 2015, pp. 1–7.

[226] A. Maxhuni, A. Munoz-Melendez, V. Osmani, H. Perez, O. Mayora, and E. F. Morales, "Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients," *ACM Pervasive and Mobile Computing*, Vol. 31, pp. 50–66, 2016.

[227] M. Faurholt-Jepsen, J. Busk, M. Frost, M. Vinberg, E. M. Christensen, O. Winther, J. E. Bardram, and L. V. Kessing, "Voice analysis as an objective state marker in bipolar disorder," *Translational Psychiatry*, Vol. 6, pp. 1–8, 2016.

[228] K. Matsumoto, G. T. Samson, O. D. O'Daly, D. K. Tracy, A. D. Patel, and S. S. Shergill, "Prosodic discrimination in patients with schizophrenia," *British Journal of Psychiatry*, Vol. 189, No. 8, pp. 180–181, 2006.

[229] M. Alpert, S. D. Rosenberg, E. R. Pouget, and R. J. Shaw, "Prosody and lexical accuracy in flat affect schizophrenia," *Psychiatry Research*, Vol. 97, No. 2-3, pp. 107–118, 2000.

[230] J. J. Diehl and R. Paul, "Acoustic Differences In The Imitation Of Prosodic Patterns In Children With Autism Spectrum Disorders." *Research in autism spectrum disorders*, Vol. 6, No. 1, pp. 123–134, 2012.

[231] R. Fusaroli and K. Tylen, "Investigating Conversational Dynamics: Interactive Alignment, Interpersonal Synergy, and Collective Task Performance," *Cognitive Science*, Vol. 40, No. 1, pp. 145–171, 2016.

[232] J. Laver, "The Phonetic Description of Voice Quality," *New York*, Vol. 31, pp. 66–92, 1980.

[233] C. Gobl and A. Ni Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Communication*, Vol. 40, No. 1-2, pp. 189–212, 2003.

[234] J. Kane and C. Gobl, "Wavelet maxima dispersion for breathy to tense voice discrimination," *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 21, No. 6, pp. 1170–1179, 2013.

[235] A. Ozdas, R. G. Shiavi, S. E. Silverman, M. K. Silverman, and D. M. Wilkes, "Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk," *IEEE Transactions on Biomedical Engineering*, Vol. 51, No. 9, pp. 1530–1540, 2004.

[236] Z. Yu, S. Scherer, D. Devault, J. Gratch, G. Stratou, L.-P. Morency, and J. Cassell, "Multimodal Prediction of Psychological Disorders: Learning Verbal and Nonverbal Commonalities in Adjacency Pairs," in *Workshop on the Semantics and Pragmatics of Dialogue*, 2013, pp. 160–169.

[237] V. Venek, S. Scherer, L. P. Morency, A. S. Rizzo, and J. Pestian, "Adolescent Suicidal Risk Assessment in Clinician-Patient Interaction," *IEEE Transactions on Affective Computing*, Vol. 8, No. 2, pp. 204–215, 2017.

[238] S. Bhatia, M. Hayat, M. Breakspear, G. Parker, and R. Goecke, "A Video-Based Facial Behaviour Analysis Approach to Melancholia," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2017, pp. 754–761.

[239] A. Guidi, J. Schoentgen, G. Bertschy, C. Gentili, L. Landini, E. P. Scilingo, and N. Vanello, "Voice quality in patients suffering from bipolar disease," in *IEEE Engineering in Medicine and Biology Society Conference (EMBC)*, 2015, pp. 6106–6109.

[240] P. Alku, H. Strik, and E. Vilkman, "Parabolic spectral parameter – A new method for quantification of the glottal flow," *Speech Communication*, Vol. 22, No. 1, pp. 67–79, 1997.

[241] J. Kane and C. Gobl, "Identifying regions of non-modal phonation using features of the wavelet transform," in *INTERSPEECH*, 2011, pp. 177–180.

[242] P. Alku, T. Backstrom, and E. Vilkman, "Normalized amplitude quotient for parametrization of the glottal flow." *The Journal of the Acoustical Society of America*, Vol. 112, No. 2, pp. 701–710, 2002.

[243] R. Fraile and J. I. Godino-Llorente, "Cepstral peak prominence: A comprehensive analysis," *Biomedical Signal Processing and Control*, Vol. 14, pp. 42–54, 2014.

[244] Y. Shue, G. Chen, and A. Alwan, "On the Interdependencies between Voice Quality, Glottal Gaps, and Voice-Source related Acoustic Measures," in *INTERSPEECH*, 2010, pp. 34–37.

[245] A. J. Flint, S. E. Black, I. Campbell-Taylor, G. F. Gailey, and C. Levinton, "Abnormal speech articulation, psychomotor retardation, and subcortical dysfunction in major depression," *Journal of Psychiatric Research*, Vol. 27, No. 3, pp. 309–319, 1993.

[246] N. Cummins, J. Epps, V. Sethu, and J. Krajewski, "Weighted pairwise Gaussian likelihood regression for depression score prediction," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4779–4783.

[247] S. Umesh, L. Cohen, and D. Nelson, "Fitting the Mel scale," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. 1, No. 1, 1999, pp. 217 –220.

[248] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *The Journal of the Acoustical Society of America*, Vol. 87, No. 4, pp. 1738–1752, 1990.

[249] F. Honig, G. Stemmer, C. Hacker, and F. Brugnara, "Revising Perceptual Linear Prediction (PLP)," in *INTERSPEECH*, 2005, pp. 2997–3000.

[250] H. Misra, S. Ikbal, H. Bourlard, and H. Herman, "Spectral entropy based feature for robust ASR," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2004, pp. 193–196.

[251] T. Leino, "Long-Term Average Spectrum in Screening of Voice Quality in Speech: Untrained Male University Students," *Journal of Voice*, Vol. 23, No. 6, pp. 671–676, 2009.

[252] B. Hammarberg, B. Fritzell, J. Gaufin, J. Sundberg, and L. Wedin, "Perceptual and acoustic correlates of abnormal voice qualities," *Acta Oto-Laryngologica*, Vol. 90, No. 1-6, pp. 441–451, 1980.

[253] P. C. Khoa and C. E. Siong, "Spectral local harmonicity feature for voice activity detection," in *International Conference on Audio, Language and Image Processing*, 2012, pp. 407–413.

[254] N. Cummins, J. Epps, V. Sethu, M. Breakspear, and R. Goecke, "Modeling spectral variability for the classification of depressed speech," in *INTERSPEECH*, 2013, pp. 857–861.

[255] P. Lopez-Otero, L. Docio-Fernandez, and C. Garcia-Mateo, "A study of acoustic features for the classification of depressed speech," in *International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2014, pp. 1331–1335.

[256] ——, "Assessing speaker independence on a speech-based depression level estimation system," *Pattern Recognition Letters*, Vol. 68, Part 2, pp. 343–350, 2015.

[257] M. Karafiat, L. Burget, P. Matejka, O. Glembek, and J. Cernocky, "iVector-based discriminative adaptation for automatic speech recognition," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011, pp. 152–157.

[258] H. Sagha, J. Deng, and B. Schuller, "The effect of personality trait, age, and gender on the performance of automatic speech valence recognition," in *ACM International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2017, pp. 1–5.

[259] S. Sandoval, V. Berisha, R. L. Utianski, J. M. Liss, and A. Spanias, "Automatic assessment of vowel space area," *The Journal of the Acoustical Society of America*, Vol. 134, No. 5, pp. 477–483, 2013.

[260] D. J. France and R. G. Shiavi, "Acoustical properties of speech as indicators of depression and suicidal risk," *IEEE Transactions on Biomedical Engineering*, Vol. 47, No. 7, pp. 829–837, 2000.

[261] J. Zhang, Z. Pan, C. Gui, J. Zhu, and D. Cui, "Clinical investigation of speech signal features among patients with schizophrenia," *Shanghai Archives of Psychiatry*, Vol. 28, No. 2, pp. 95–102, 2016.

[262] Y. Bea, T. F. Quatieri, J. R. Williamson, and J. C. Mundt, "Prediction of cognitive performance in an animal fluency task based on rate and articulatory markers," in *INTERSPEECH*, 2014, pp. 1038–1042.

[263] S. Motlagh, H. Moradi, and H. Pouretemad, "Using general sound descriptors for early autism detection," in *IEEE Asian Control Conference (ASCC)*, 2013, pp. 1–5.

[264] A. N. Finnerty, S. Muralidhar, L. S. Nguyen, F. Pianesi, and D. Gatica-Perez, "Stressful first impressions in job interviews," in *ACM International Conference on Multimodal Interaction (ICMI)*, 2016, pp. 325–332.

[265] D. McDuff, J. M. Girard, and R. el Kaliouby, "Large-Scale Observational Evidence of Cross-Cultural Differences in Facial Behavior," *Journal of Nonverbal Behavior*, Vol. 41, No. 1, pp. 1–19, 2017.

[266] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Muller, and S. Narayanan, "Paralinguistics in speech and language — State-of-the-art and the challenge," *Computer Speech and Language*, Vol. 27, No. 1, pp. 4–39, 2013.

[267] A. M. Kring and A. H. Gordon, "Sex differences in emotion: Expression, experience, and physiology." *Journal of Personality and Social Psychology*, Vol. 74, No. 3, pp. 686–703, 1998.

[268] J. J. Gross, O. P. John, and J. M. Richards, "The dissociation of emotion expression from emotion experience: A personality perspective," *Personality and Social Psychology Bulletin*, Vol. 26, No. 6, pp. 712–726, 2000.

[269] C. J. Price, "A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading," *NeuroImage*, Vol. 62, No. 2, pp. 816–847, 2012.

[270] D. Bone, T. Chaspari, K. Audhkhasi, J. Gibson, A. Tsiartas, M. Van Segbroeck, M. Li, S. Lee, and S. S. Narayanan, "Classifying Language-Related Developmental Disorders from Speech Cues: the Promise and the Potential Confounds," in *INTERSPEECH*, 2013, pp. 182–186.

[271] A. Karasz, "Cultural differences in conceptual models of depression," *Social Science & Medicine*, Vol. 60, pp. 1625–1635, 2005.

[272] N. Ambady and R. Rosenthal, "Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis," *Psychological Bulletin*, Vol. 111, No. 2, pp. 256–274, 1992.

[273] N. Ambady, F. J. Bernieri, and J. A. Richeson, "Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream," *Advances in Experimental Social Psychology*, Vol. 32, No. 0065-2601, pp. 201–271.

[274] L. S. Nguyen and D. Gatica-Perez, "I Would Hire You in a Minute: Thin Slices of Nonverbal Behavior in Job Interviews," in *ACM on International Conference on Multimodal Interaction (ICMI)*, 2015, pp. 51–58.

[275] N. Jaques, D. McDuff, Y. L. Kim, and R. Picard, "Understanding and Predicting bonding in conversations using thin slices of facial expressions and body language," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 10011 LNAI, 2016, pp. 64–74.

[276] G. Kiss, J. P. van Santen, E. Prud'hommeaux, and L. M. Black, "Quantitative analysis of pitch in speech of children with neurodevelopmental disorders," in *INTERSPEECH*, Vol. 2, 2012, pp. 1342–1345.

[277] A. S. Cohen, T. L. Renshaw, K. R. Mitchell, and Y. Kim, "A psychometric investigation of "macroscopic" speech measures for clinical and psychological science," *Behavior Research Methods*, Vol. 48, No. 2, pp. 475–486, 2016.

[278] World Health Organisation, "Depression fact sheet," 2017. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs369/en/

[279] J. Hamalainen, K. Poikolainen, E. Isometsa, J. Kaprio, M. Heikkinen, S. Lindeman, and H. Aro, "Major depressive episode related to long unemployment and frequent alcohol intoxication." *Nordic journal of psychiatry*, Vol. 59, No. 6, pp. 486–91, 2005.

[280] L. V. Kessing, "Depression and the risk for dementia," *Current Opinion in Psychiatry*, Vol. 25, No. 6, pp. 457–461, 2012.

[281] P. Gorwood, S. Richard-Devantoy, F. Bayle, and M. L. Clery-Melun, "Psychomotor retardation is a scar of past depressive episodes, revealed by simple cognitive tests," *European Neuropsychopharmacology*, Vol. 24, No. 10, pp. 1630–1640, 2014.

[282] Stanford Mood and Anxiety Disorders Laboratory, "Symptoms of Depression," 2018. [Online]. Available: https://web.stanford.edu/group/mood/Pages/AboutSymptoms.html

[283] P. M. Niedenthal, L. W. Barsalou, P. Winkielman, S. Krauth-Gruber, and F. Ric, "Embodiment in Attitudes, Social Perception, and Emotion," *Personality and Social Psychology Review*, Vol. 9, No. 3, pp. 184–211, 2005.

[284] D. Schrijvers, W. Hulstijn, and B. G. C. Sabbe, "Psychomotor symptoms in depression: A diagnostic, pathophysiological and therapeutic tool," *Journal of Affective Disorders*, Vol. 109, No. 1-2, pp. 1–20, 2008.

[285] A. J. Rush, M. H. Trivedi, H. M. Ibrahim, T. J. Carmody, B. Arnow, D. N. Klein, J. C. Markowitz, P. T. Ninan, S. Kornstein, R. Manber, M. E. Thase, J. H. Kocsis, and M. B. Keller, "The 16-item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): A psychometric evaluation in patients with chronic major depression," *Biological Psychiatry*, Vol. 54, No. 5, pp. 573–583, 2003.

[286] C. Cusin, H. Yang, A. Yeung, and M. Fava, "Rating Scales for Depression," in *Handbook of Clinical Rating Scales and Assessment in 7 Psychiatry and Mental Health*, 2010, pp. 487–488.

[287] E. I. Fried, S. Epskamp, R. M. Nesse, F. Tuerlinckx, and D. Borsboom, "What are 'good' depression symptoms? Comparing the centrality of DSM and non-DSM symptoms of depression in a network analysis," *Journal of Affective Disorders*, Vol. 189, No. Supplement C, pp. 314–320, 2016.

[288] E. I. Fried, "Are more responsive depression scales really superior depression scales?" *Journal of Clinical Epidemiology*, Vol. 77, pp. 4–6, 2016.

[289] A. T. Beck, R. A. Steer, R. Ball, and W. Ranieri, "Comparison of Beck Depression Inventories - IA and -II in psychiatric outpatients," *Journal of Personality Assessment*, Vol. 67, No. 3, pp. 588–597, 1996.

[290] D. M. W. Powers, "Evaluation: From Precision, Recall and F-Measure To Roc, Informedness, Markedness & Correlation," *Journal of Machine Learning Technologies*, Vol. 2, No. 1, pp. 37–63, 2011.

[291] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 7, pp. 971–987, 2002.

[292] Jianbo Shi and Tomasi, "Good features to track," in *IEEE Computer Vision and Pattern Recognition (CVPR)*, 1994, pp. 593–600.

[293] M. V. Segbroeck, A. Tsiartas, and S. S. Narayanan, "A robust frontend for VAD: Exploiting contextual, discriminative and spectral cues of human voice," in *INTERSPEECH*, 2013, pp. 704–708.

[294] M. Kachele, M. Schels, and F. Schwenker, "Inferring Depression and Affect from Application Dependent Meta Knowledge," in *ACM International Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2014, pp. 41–48.

[295] H. Perez Espinosa, H. J. Escalante, L. Villasenor-Pineda, M. Montes-y Gomez, D. Pinto-Avedano, and V. Reyez-Meza, "Fusing Affective Dimensions and Audio-Visual Features from Segmented Video for Depression Recognition: INAOE-BUAP's Participation at AVEC'14 Challenge," in *ACM International Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2014, pp. 49–55.

[296] J. Nelder and R. Wedderburn, "Generalized Linear Models," *Journal of the Royal Statistical Society: Series A*, Vol. 135, pp. 370–384, 1972.

[297] V. Jain, J. L. Crowley, A. K. Dey, and A. Lux, "Depression Estimation Using Audiovisual Features and Fisher Vector Encoding," in *ACM International Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2014, pp. 87–91.

[298] H. Wang, A. Klaser, C. Schmid, and C. L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International Journal of Computer Vision*, Vol. 103, No. 1, pp. 60–79, 2013.

[299] A. Jan, H. Meng, Y. F. A. Gaus, F. Zhang, and S. Turabzadeh, "Automatic Depression Scale Prediction using Facial Expression Dynamics and Regression," in *ACM International Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2014, pp. 1–7.

[300] H. Meng, N. Pears, M. Freeman, and C. Bailey, "Motion History Histograms for Human Action Recognition," in *Embedded Computer Vision SE - 7*, ser. Advances in Pattern Recognition, B. Kisacanin, S. Bhattacharyya, and S. Chai, Eds. Springer London, 2009, pp. 139–162.

[301] V. Ojansivu and J. Heikkila, "Blur insensitive texture classification using local phase quantization," in *Lecture Notes in Computer Science*, Vol. 5099, 2008, pp. 236–243.

[302] J. R. Williamson, D. W. Bliss, and D. W. Browne, "Epileptic seizure prediction using the spatiotemporal correlation structure of intracranial EEG," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 665–668.

[303] J. R. Williamson, A. Dumas, G. Ciccarelli, A. R. Hess, B. A. Telfer, and M. J. Buller, "Estimating load carriage from a body-worn accelerometer," in *IEEE International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, 2015, pp. 1–6.

[304] H. Meng, D. Huang, H. Wang, H. Yang, M. AI-Shuraifi, and Y. Wang, "Depression Recognition Based on Dynamic Facial and Vocal Expression Features Using Partial Least Square Regression," in *ACM International Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2013, pp. 21–30.

[305] X. Ma, H. Yang, Q. Chen, D. Huang, and Y. Wang, "DepAudioNet: An Efficient Deep Model for Audio Based Depression Classification," in *ACM International Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2016, pp. 35–42.

[306] L. Yang, D. Jiang, X. Xia, E. Pei, M. C. Oveneke, and H. Sahli, "Multimodal Measurement of Depression Using Deep Learning Models," in *ACM International Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2017, pp. 1–7.

[307] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International Conference on International Conference on Machine Learning*, 2014, pp. 1188–1196.

[308] T. K. Witte, K. A. Timmons, E. Fink, A. R. Smith, and T. E. Joiner, "Do major depressive disorder and dysthymic disorder confer differential risk for suicide?" *Journal of Affective Disorders*, Vol. 115, No. 1-2, pp. 69–78, 2009.

[309] K. Kyle, "Suite of Automatic Linguistic Analysis Tools (SALAT)," 2018. [Online]. Available: http://www.kristopherkyle.com/

[310] M. M. Bradley and P. J. Lang, "Affective norms for English words (ANEW): Instruction manual and affective ratings," Technical report C-1 , the Center for Research in Psychophysiology, University of Florida, Tech. Rep., 1999.

[311] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The Development and Psychometric Properties of LIWC 2015," *Austin, TX: University of Texas at Austin*, pp. 1–22, 2015.

[312] M. Hall, "Correlation-based Feature Selection for Machine Learning," *Methodology*, No. 4, pp. 1–5, 1999.

[313] W. M. Hartmann and J. Pumplin, "Periodic signals with minimal power fluctuations." *The Journal of the Acoustical Society of America*, Vol. 90, No. 4, pp. 1986–1999, 1991.

[314] D. Ebert, R. Albert, G. Hammon, B. Strasser, A. May, and A. Merz, "Eye-blink rates and depression. Is the antidepressant effect of sleep deprivation mediated by the dopamine system?" *Neuropsychopharmacology*, Vol. 15, No. 4, pp. 332–339, 1996.

[315] H. Kaya and A. A. Salah, "Eyes Whisper Depression," in *ACM International Conference on Multimedia*, 2014, pp. 961–964.

[316] K. Simonyan, O. Parkhi, A. Vedaldi, and A. Zisserman, "Fisher Vector Faces in the Wild," in *British Machine Vision Conference*, 2013, pp. 1–12.

[317] H. Kaya, F. Gurpinar, S. Afshar, and A. A. Salah, "Contrasting and Combining Least Squares Based Learners for Emotion Recognition in the Wild," in *ACM International Conference on Multimodal Interaction (ICMI)*, 2015, pp. 459–466.

[318] H. Kaya and A. A. Karpov, "Fusing Acoustic Feature Representations for Computational Paralinguistics Tasks," in *INTERSPEECH*, 2016, pp. 2046–2050.

[319] T. S. Jaakkola and D. Haussler, "Exploiting generative models in discriminative classifiers," in *Advances in Neural Information Processing Systems*, 1998, pp. 487–493.

[320] H. Jegou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid, "Aggregating local image descriptors into compact codes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 34, No. 9, pp. 1704–1716, 2012.

[321] A. Vedaldi and B. Fulkerson, "VLFeat: An Open and Portable Library of Computer Vision Algorithms," in *ACM International Conference on Multimedia*, 2010, pp. 1469–1472.

[322] W. Zong, G. B. Huang, and Y. Chen, "Weighted extreme learning machine for imbalance learning," *Neurocomputing*, Vol. 101, pp. 229–242, 2013.

[323] R. Rummer, J. Schweppe, R. Schlegelmilch, and M. Grice, "Mood is linked to vowel type: The role of articulatory movements," *Emotion*, Vol. 14, No. 2, pp. 246–250, 2014.

[324] S. Scherer, G. M. Lucas, J. Gratch, A. Rizzo, and L. P. Morency, "Self-Reported Symptoms of Depression and PTSD Are Associated with Reduced Vowel Space in Screening Interviews," *IEEE Transactions on Affective Computing*, Vol. 7, No. 1, pp. 59–73, 2016.

[325] Word Health Organisation, "Word Health Organisation Factsheets — Mental Disorders," 2018. [Online]. Available: http://www.who.int/news-room/fact-sheets/detail/mental-disorders

[326] E. Severus and M. Bauer, "Diagnosing bipolar disorders in DSM-5," *International Journal of Bipolar Disorders*, Vol. 1, No. 14, pp. 1–3, 2013.

[327] I. Grande, M. Berk, B. Birmaher, and E. Vieta, "Bipolar disorder," *The Lancet*, Vol. 387, No. 10027, pp. 1561–1572, 2016.

[328] L. Deng and D. O'Shaughnessy, *Speech Processing: A Dynamic and Optimization-Oriented Approach*, 1st ed. CRC Press, 2003.

[329] S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, M. Christian, S. Language, P. Group, D. Telekom, and A. G. Laboratories, "The INTERSPEECH 2010 Paralinguistic Challenge," in *INTERSPEECH*, 2010, pp. 2794–2797.

[330] ACII-ASIA, "Asian Conference on Affective Computing and Intelligent Interaction,," 2018. [Online]. Available: www.acii-asia-2018.org

[331] R. Schleicher, N. Galley, S. Briest, and L. Galley, "Blinks and saccades as indicators of fatigue in sleepiness warnings: Looking tired?" *Ergonomics*, Vol. 51, No. 7, pp. 982–1010, 2008.

[332] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer, "On the acoustics of emotion in audio: What speech, music, and sound have in common," *Frontiers in Psychology*, Vol. 4, No. 5, pp. 1–12, 2013.

[333] P. Guo, P. C. L. Chen, and Y. Sun, "An Exact Supervised Learning for a Three-Layer Supervised Neural Network," in *International Conference on Neural Information Processing*, 1995, pp. 1041–1044.

[334] W. B. Johnson and J. Lindenstrauss, "Extensions of Lipschitz mappings into a Hilbert space," in *Conference in modern analysis and probability (New Haven, Conn., 1982) Contemporary Mathematics. 26. Providence, RI: American Mathematical Society*, 1984, pp. 189–206.

[335] T. M. T. Cover, "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition," *IEEE Transactions on Electronic Computers*, Vol. 14, No. 3, pp. 326–334, 1965.

[336] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, and B. W. Schuller, "An Image-based Deep Spectrum Feature Representation for the Recognition of Emotional Speech," in *ACM on Multimedia Conference*, 2017, pp. 478–484.

[337] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, and M. Pantic, "Summary for AVEC 2018: Bipolar Disorder and Cross-Cultural Affect Recognition," in *Audio/Visual Emotion Challenge Workshop*, 2018, pp. 2111–2112.

[338] Z. Du, W. Li, D. Huang, and Y. Wang, "Bipolar Disorder Recognition via Multi-scale Discriminative Audio Temporal Representation," in *Audio/Visual Emotion Challenge Workshop*, 2018, pp. 23–30.

[339] X. Xing, B. Cai, Y. Zhao, S. Li, Z. He, and W. Fan, "Multi-modality Hierarchical Recall based on GBDTs for Bipolar Disorder Classification," in *ACM Audio/Visual Emotion Challenge Workshop*, 2018, pp. 31–37.

[340] L. Yang, Y. Li, H. Chen, D. Jiang, M. C. Oveneke, and H. Sahli, "Bipolar Disorder Recognition with Histogram Features of Arousal and Body Gestures," in *Audio/Visual Emotion Challenge Workshop*, 2018, pp. 15–21.

[341] C. Szegedy, V. Vanhoucke, S. Ioffe, and J. Shlens, "Rethinking the Inception Architecture for Computer Vision," in *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826.

[342] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning," in *AAAI Conference on Artifcial Intelligence*, 2017, pp. 4278–4284.

[343] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, Vol. 9, No. 8, pp. 1735–1780, 1997.

[344] Megvii Technology Co Ltd., "Face++." [Online]. Available: https://www.faceplusplus.com.cn/

[345] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[346] L. He, D. Jiang, L. Yang, E. Pei, P. Wu, and H. Sahli, "Multimodal Affective Dimension Prediction Using Deep Bidirectional Long Short-Term Memory Recurrent Neural Networks," in *Audio/Visual Emotion Challenge Workshop*, 2015, pp. 73–80.

[347] F. Ringeval, B. Schuller, S. Jaiswal, M. Valstar, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic, "AV+EC 2015 – The First Affect Recognition Challenge Bridging Across Audio, Video, and Physiological Data," in *Proceedings of the 5th International Workshop on AVEC, ACM MM*, 2015, pp. 3–8.

[348] Z. Cao, T. Simon, S. E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 7291–7299.

[349] P. Pudil, J. Novovicova, and J. Kittler, "Floating search methods in feature selection," *Pattern Recognition Letters*, Vol. 15, No. 11, pp. 1119–1125, 1994.

[350] D. Ververidis and C. Kotropoulos, "Fast sequential floating forward selection applied to emotional speech features estimated on des and SUSAS data collections," in *European Signal Processing Conference (EUSIPCO)*, 2006, pp. 2219–5491.

[351] Z. S. Syed, K. Sidorov, and D. Marshall, "Automated Screening for Bipolar Disorder from Audio/Visual Modalities," in *ACM International Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2018, pp. 39–45.

[352] M. Sipser, *Introduction to the Theory of Computation*, 3rd ed. Cengage Learning, 1996.

[353] S. Salvador and P. Chan, "Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms," in *International Conference on Tools with Artificial Intelligence*, 2004.

[354] H. Akaike, "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control*, Vol. 19, No. 6, pp. 716–723, 1974.

[355] World Health Organisation, "World Health Organisation Factsheets — Autism Spectrum Disorders," 2018. [Online]. Available: http://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders

[356] Centers for Disease Control and Prevention, "Autism Spectrum Disorder: Statistics," 2018. [Online]. Available: https://www.cdc.gov/ncbddd/autism/data.html

[357] The National Autistic Society UK, "Autism Statistics for United Kingdom," 2018. [Online]. Available: http://www.autism.org.uk/about/what-is/myths-facts-stats.aspx

[358] Y. Brukner-Wertman, N. Laor, and O. Golan, "Social (Pragmatic) Communication Disorder and Its Relation to the Autism Spectrum: Dilemmas Arising From the DSM-5 Classification," *Journal of Autism and Developmental Disorders*, Vol. 46, No. 8, pp. 2821–2829, 2016.

[359] A. Metallinou, R. B. Grossman, and S. Narayanan, "Quantifying atypicality in Affective Facial Expressions of Children with Autism Spectrum Disorders," in *IEEE International Conference on Multimedia and Expo*, 2013, pp. 1–6.

[360] W. Liu, L. Yi, Z. Yu, X. Zou, B. Raj, and M. Li, "Efficient autism spectrum disorder prediction with eye movement: A machine learning framework," in *International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2015, pp. 649–655.

[361] J. M. Rehg, "Behavior Imaging: Using Computer Vision to Study Autism," in *IAPR Conference on Machine Vision Applications*, 2011, pp. 14–21.

[362] S. S. Rajagopalan, A. Dhall, and R. Goecke, "Self-Stimulatory Behaviours in the Wild for Autism Diagnosis," in *Computer Vision Workshops (ICCVW), 2013*, 2013, pp. 755–761.

[363] J. M. Rehg, A. Rozga, G. D. Abowd, and M. S. Goodwin, "Behavioral Imaging and Autism," *IEEE Pervasive Computing*, Vol. 13, No. 2, pp. 84–87, 2014.

[364] N. Goncalves, S. Costa, J. Rodrigues, and F. Soares, "Detection of stereotyped hand flapping movements in Autistic children using the Kinect sensor: A case study," in *Autonomous Robot Systems and Competitions (ICARSC)*, 2014, pp. 212–216.

[365] M. Sharda, T. P. Subhadra, S. Sahay, C. Nagaraja, L. Singh, R. Mishra, A. Sen, N. Singhal, D. Erickson, and N. C. Singh, "Sounds of melody-Pitch patterns of speech in autism," *Neuroscience Letters*, Vol. 478, No. 1, pp. 42–45, 2010.

[366] Y. S. Bonneh, Y. Levanon, O. Dean-Pardo, L. Lossos, and Y. Adini, "Abnormal Speech Spectrum and Increased Pitch Variability in Young

Autistic Children," *Frontiers in Human Neuroscience*, Vol. 4, pp. 1–7, 2010.

[367] Y. Kakihara, T. Takiguchi, Y. Ariki, Y. Nakai, and S. Takada, "Acoustic feature selection utilizing multiple kernel learning for classification of children with autism spectrum and typically developing children," in *IEEE/SICE International Symposium on System Integration (SII)*, 2013, pp. 490–494.

[368] A. V. Ivanov, S. Jalalvand, R. Gretter, and D. Falavigna, "Phonetic and anthropometric conditioning of MSA-KST cognitive impairment characterization system," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, 2013, pp. 228–233.

[369] J. McCann and S. Peppe, "Prosody in autism spectrum disorders: A critical review," *International Journal of Language and Communication Disorders*, Vol. 38, pp. 325–350, 2003.

[370] P. Mokhtari and N. Campbell, "Voice Quality: the 4th Prosodic Dimension," in *International Congress of Phonetic Sciences*, 2003, pp. 2417–2420.

[371] T. Johnstone and K. R. Scherer, "The Effects of Emotions on Voice Quality," *International Conference of Phonetic Sciences*, pp. 2029–2032, 1999.

[372] I. Grichkovtsova, M. Morel, and A. Lacheret, "The role of voice quality and prosodic contour in affective speech perception," *Speech Communication*, Vol. 54, No. 3, pp. 414–429, 2012.

[373] K. Kirchhoff, Y. Liu, and J. Bilmes, "Classification of developmental disorders from speech signals using submodular feature selection," in *INTERSPEECH*, 2013, pp. 187–190.

[374] O. Rasanen and J. Pohjalainen, "Random subset feature selection in automatic recognition of developmental disorders, affective states, and level of conflict from speech," in *INTERSPEECH*, 2013, pp. 210–214.

[375] D. Martinez, D. Ribas, E. Lleida, A. Ortega, and A. Miguel, "Suprasegmental information modelling for autism disorder spectrum and specific language impairment classification," in *INTERSPEECH*, 2013, pp. 195–199.

[376] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, and P. A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech and Language*, Vol. 20, No. 2, pp. 210–229, 2006.

[377] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *INTERSPEECH*, 2009, pp. 312–315.

[378] M. A. Bautista, A. Hernandez, S. Escalera, L. Igual, O. Pujol, J. Moya, V. Violant, and M. T. Anguera, "A Gesture Recognition System for Detecting Behavioral Patterns of ADHD," *IEEE Transactions on System, Man and Cybernetics, Part B*, Vol. 46, No. 1, pp. 136–147, 2015.

[379] British Council, "Voiced and unvoiced consonants," 2018. [Online]. Available: https://www.teachingenglish.org.uk/article/voiced-unvoiced-consonants-0

[380] T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *INTERSPEECH*, 2011, pp. 1973–1976.

[381] T. Drugman, Y. Stylianou, Y. Kida, and M. Akamine, "Voice Activity Detection: Merging Source and Filter-based Information," *IEEE Signal Processing Letters*, Vol. 23, No. 2, pp. 252–256, 2016.

[382] A. I. Koutrouvelis, G. P. Kafentzis, N. D. Gaubitch, and R. Heusdens, "A fast method for high-resolution voiced/unvoiced detection and glottal closure/opening instant estimation of speech," *IEEE/ACM Transactions on Audio Speech and Language Processing*, Vol. 24, No. 2, pp. 316–328, 2016.

[383] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, Vol. 6, No. 1, pp. 1–3, 1999.

[384] M. Brookes, "Voicebox," 2018. [Online]. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[385] P. Drotar, J. Gazda, and Z. Smekal, "An experimental comparison of feature selection methods on two-class biomedical datasets," *Computers in Biology and Medicine*, Vol. 66, pp. 1–10, 2015.

[386] Wolfram, "Bonferroni Correction," 2018. [Online]. Available: http://mathworld.wolfram.com/BonferroniCorrection.html

[387] P. Laukka, D. Neiberg, M. Forsell, I. Karlsson, and K. Elenius, "Expression of affect in spontaneous speech: Acoustic correlates and automatic detection of irritation and resignation," *Computer Speech and Language*, Vol. 25, No. 1, pp. 84–104, 2011.

[388] F. Ringeval, J. Demouy, G. Szaszak, M. Chetouani, L. Robel, J. Xavier, D. Cohen, and M. Plaza, "Automatic intonation recognition for the prosodic assessment of language-impaired children," *IEEE Transactions*

on Audio, Speech and Language Processing, Vol. 19, No. 5, pp. 1328–1342, 2011.

[389] E. Lyakso, O. Frolova, and A. Grigorev, "Perception and Acoustic Features of Speech of Children with Autism Spectrum Disorders," in *International Conference on Speech and Computer*, 2017, pp. 602–612.

[390] S. O. Sadjadi and J. H. Hansen, "Unsupervised speech activity detection using voicing measures and perceptual spectral flux," *IEEE Signal Processing Letters*, Vol. 20, No. 3, pp. 197–200, 2013.

[391] L. Kanner, "Autistic disturbances of affective contact," *Nervous Child*, Vol. 2, pp. 217–250, 1943.

[392] C. Kaland, E. Krahmer, and M. Swerts, "Contrastive intonation in autism: The effect of speaker-and listener-perspective." in *INTERSPEECH*, 2012, pp. 1–5.

[393] M. Tahon and L. Devillers, "Towards a small set of robust acoustic features for emotion recognition: challenges," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, Vol. 24, No. 1, pp. 16–28, 2016.

[394] S. Scherer, J. Pestian, and L. P. Morency, "Investigating the speech characteristics of suicidal adolescents," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 709–713.