5-2018

Multimodal Depression Detection: An Investigation of Features and Fusion Techniques for Automated Systems

Michelle Renee Morales
The Graduate Center, City University of New York

Multimodal Depression Detection:
An investigation of features and fusion techniques for automated systems

by

Michelle Renee Morales

A dissertation submitted to the Graduate Faculty in Linguistics in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York

2018

Multimodal Depression Detection:
An Investigation of Features and Fusion Techniques for Automated Systems

by

Michelle Renee Morales

This manuscript has been read and accepted by the Graduate Faculty in Linguistics in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

Date

Rivka Levitan

Chair of Examining Committee

Date

Gita Martohardjono

Executive Officer

Supervisory Committee:

Martin Chodorow

Andrew Rosenberg

Stefan Scherer

THE CITY UNIVERSITY OF NEW YORK

Abstract

Multimodal Depression Detection:
An Investigation of Features and Fusion Techniques for Automated Systems

by

Michelle Renee Morales

Advisor: Rivka Levitan

Depression is a serious illness that affects a large portion of the world's population. Given the large effect it has on society, it is evident that depression is a serious health issue. This thesis evaluates, at length, how technology may aid in assessing depression. We present an in-depth investigation of features and fusion techniques for depression detection systems. We also present OpenMM: a novel tool for multimodal feature extraction. Lastly, we present novel techniques for multimodal fusion. The contributions of this work add considerably to our knowledge of depression detection systems and have the potential to improve future systems by incorporating that knowledge into their design.

Acknowledgments

I would like to express my deepest gratitude to all the people who have supported, encouraged, and guided me through this long but fulfilling path in life.

To my advisor, Rivka Levitan, I express my warmest gratitude. Over the years, I have benefited enormously from your knowledge, support, and advice. Your course in *Natural Language Processing and Psychology* completely changed the trajectory of my academic path; it help me discover a passion that led to my thesis topic. For that, I am forever grateful.

Exceptional thanks go to my former advisor, Andrew Rosenberg. I will always be envious of IBM for stealing you away in my final years. Thank you for giving me a chance. Your enormous support gave me the strength to bravely pursue research that challenged my knowledge and abilities. By welcoming me into the Speech Lab, you gave me a place to learn from the smartest people. It is the place that I learned and grew most.

I cannot express my gratitude enough to the other members of my committee, Martin Chodorow and Stefan Scherer. I truly appreciate the learning opportunities, guidance, and advice that you both provided at different times in different capacities.

I would also like to gratefully acknowledge my funding from the Graduate Center's Presidential MAGNET fellowship program and the Futures Initiative program. Without the support of these programs the realization of this work would not have been possible. I would especially like to thank the leaders of these programs —Herman Bennett, Eric Frankson, Cathy Davidson, Katina Rogers —for their endless guidance and support.

A special thank you to Rachel Rakov, Corbin Neuhauser, Christen Madsen, and David Guy Brizan for your friendship. It has been a pleasure sharing my graduate student experience with you all.

To my CFGC community, thank you for giving me a healthy outlet for my stress and frustration. You have made me stronger both mentally and physically.

To my family, I can not express in words how immensely blessed I am to have you all in my life. To Cristina and Bernardo, thank you for always taking care of me. Your honesty, humor, and love are a constant support. To my best friend, Tim, thank you for absolutely everything. I could not have asked for a better partner in life. You make me feel like anything is possible.

Finally, I would like to thank and dedicate this thesis to my parents, Miguel and Ines Morales. You have been and will always be my role models. You came to the United States with nothing but your brilliant minds and unwavering passion to succeed. I am forever grateful for the amazing life you have built for me. You taught me the importance of education and were the first to show me how fun learning can be. Without you both, I would not be here. This thesis is a much a culmination of your hard work and sacrifice as it is mine. I love you and I thank you from the bottom of my heart.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Depression is a serious mental health issue that affects millions of people globally. In the United States, depression affects approximately 14.8 million adults, or about 6.7 percent of the U.S. population age 18 and older (Anxiety and Depression Association of America [ADAA], 2015). The World Health Organization estimates that by the year 2020, depression will be the second largest cause of burden of disease worldwide, and by the year 2030, it is expected to be the largest cause (World Health Organization [WHO], 2001). Due to the variation in how depression presents itself within each person, it is difficult and time consuming to diagnose. Since diagnosis often relies on a clinician's assessment, it is also subjective. Moreover, many under-served regions have severe shortages of clinicians who can make the diagnosis. Even in areas with well-developed health systems, less than half of those suffering from depression receive treatment (WHO, 2001).

Given advancements in hardware and software, coupled with the explosion of smart phone use, possible health care solutions have begun to change and interest in developing technologies to assess depression has grown. Most prominent among the research in depression detection, has been the use of speech as an objective marker of depression (Cummins et al., 2015a). Researchers have also investigated the potential of visual cues for assessing depression level (Scherer et al., 2014). Some have even combined audio and visual cues, building multimodal systems to assess depression level. We use modality to refer to certain types of information, representations, channels, or signals. For example, modalities include language (both spoken or written), visual (from images or videos), auditory, (including voice, sounds and music). Multimodality is then defined as the presence of more than one modality, e.g. audio, video, and language (Poria et al., 2017). Depression detection studies which have systematically compared unimodal systems with multimodal ones have consistently found multimodal systems perform best (Alghowinem et al., 2015; Joshi et al., 2013a; Morales and Levitan, 2016b; Morales et al., 2017b). These findings support past work from the more general multimodal machine learning literature, which found multimodal systems offer various advantages (Baltrušaitis et al., 2016). For example, having access to multiple modalities that observe the same phenomenon may allow for more robust predictions, as it allows for complementary information from each modality; something not visible in individual modalities may appear when using multiple ones. In addition, a multimodal system can still operate when one of the modalities is missing. Empirically, related work on multimodal affect recognition, has found that, on average, multimodal systems offer an 8% relative improvement over unimodal systems (D'Mello and Kory, 2012).

Specifically for the task of depression detection, domain knowledge strongly motivates classification based on multiple modalities. Given the extensive body of research on objective markers of depression, it is clear that depression affects an individual in many ways. Researchers have found relationships between depression and a number of markers, including biological, physiological, nonverbal, and verbal (Cummins et al., 2015b). Therefore, in order to provide a comprehensive representation of an individual's possible depression symptoms, we must provide the depression detection system with multimodal information. In recent years, researchers have noted the promise of multimodal systems. As a result, significant progress has been made, but many challenges remain. Most relevant to this work are the following challenges related to multimodal depression detection:

1. Disconnect between research fields: approaches to depression assessment from psychology, natural language processing (NLP), speech processing, and human-computer interaction (HCI) tend to silo by subfield, with little discussion about the utility of combining promising approaches
2. Lack of research on how to combine modalities (i.e. fusion techniques) for this task
3. Need for tools to facilitate greater collaboration and cooperation, especially across modalities

Investigating these challenges and presenting possible solutions represent the main contributions of this thesis. In the following section, the three main research goals of this work are described. Our work adds to the body of knowledge on multimodal depression detection systems and proposes new directions for making use of that knowledge to enhance detection systems.

1.1 Research Goals

1.1.1 Bridge the Disconnect between Research Fields

The first goal of this work is to help bridge the disconnect that exists between the depression detection research fields (Morales et al., 2017a). This research goal is the central overarching aim of this thesis and each subsequent goal contributes to this primary objective. This goal is of primary importance because in order to build a successful depression detection system that can be used in practice, there is a need for greater connections between disciplines. Each research field represents one component, which provides an important piece of understanding, that is necessary to build a comprehensive detection system, which can represent and measure

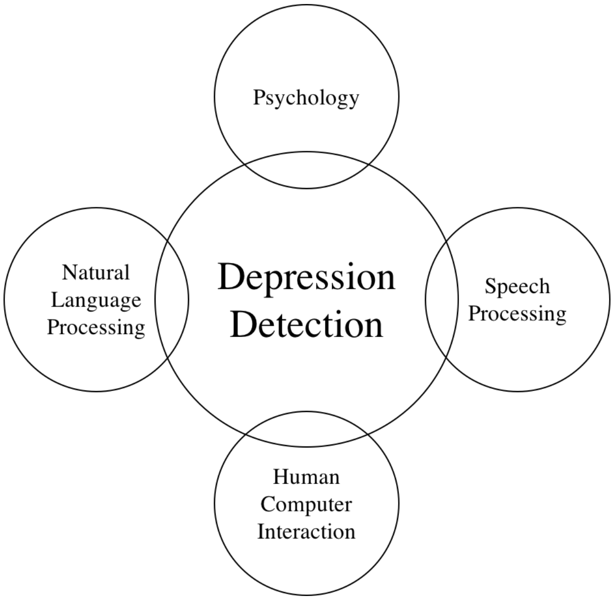


Figure 1.1: Research landscape for depression detection systems

an individual's entire behavior. Therefore, to truly build an all-inclusive and accurate system collaboration between fields is necessary.

The current research landscape is segmented and each group is siloed by discipline, as depicted in Figure 1.1. Experts across several fields are attempting to build valid tools for depression detection. Although these research fields have the same goal in mind, they approach the problem from different perspectives and the methodology used by each varies substantially; each field focuses on different modalities, uses different datasets and tools, and employs different evaluation metrics. Due to these experimental differences, it is difficult to compare approaches and even more difficult to combine promising approaches. For example, if we consider data sources alone, NLP research has aimed to detect depression from writing, both formal and informal (i.e. online text), speech processing research has aimed to assess depression level from audio while HCI and related fields try to assess depression level from video. Each data source is then labeled for depression through different approaches, including rating scales, self-report surveys, manual annotation, etc. As a result, we see various definitions of how depression is defined across studies. Given these differences, it becomes difficult to make systematic comparisons across fields. Collaboration and cooperation between fields is also rare. Regardless of the existing differences, every study and system share the common goal of discovering a way to use technology to help assess depression.

We aim to bridge the disconnect between the research fields by providing the research community with a comprehensive resource: an in-depth literature review of depression detection systems across modalities and fields. This review helps inform the entire research community about ongoing research in different fields, highlighting each approach's strengths and weaknesses. The review also provides an overview of existing datasets as well as descriptions of features and existing machine learning approaches. In addition, the review outlines existing evaluation metrics across fields to help promote uniformity. By providing a cross-modal and cross-field review of detection systems, we help connect the disparate fields through the dissemination of knowledge. In the next two sub-sections, we discuss our second and third research goals, which also aim to address our primary goal. Specifically, we introduce our work in multimodal depression detection systems as well as present our novel open-source tool, OpenMM.

1.1.2 In-depth Investigation of Multimodal Fusion Techniques

The second research goal of this work is to investigate fusion techniques for multimodal depression detection systems. The framework for building multimodal systems consists of two fundamental steps: processing unimodal data separately and fusing them all together (Poria et al., 2017). Multimodal fusion is defined as the concept of integrating information from multiple modalities with the goal of predicting an outcome measure (Baltrušaitis et al., 2016). Most research on depression detection has evaluated how well a system can predict depression score, with most efforts aimed at discovering what features lead to the best system performance. Therefore, most research has focused on unimodal data. Although some research has presented multimodal systems and evaluated fusion techniques, this work is the first to thoroughly investigate fusion techniques for multimodal depression detection. This investigation includes an investigation of popular existing fusion approaches, such as early and late fusion. In addition, this thesis presents and evaluates a novel fusion approach: *informed fusion*. Understanding fusion techniques for depression detection is critical for designing successful multimodal systems. In addition, this research also helps address our first research goal: bridging the existing disconnect. By building and evaluating unimodal (visual, acoustic, and linguistic) and multimodal systems, we present a systematic comparison of approaches from each research field. In our approach to multimodal design we borrow approaches from NLP, speech processing, and HCI. By systematically evaluating existing approaches across the various fields, we help achieve a better understanding of each approach's promise and each field's contributions. Moreover, each feature set investigated is motivated from the psychological literature and in our evaluation we aim to connect our results back to the psychology of depression. By demonstrating the promise of multimodal systems we also hope to promote more multimodal research for depression detection and in turn lessen the gap between fields.

1.1.3 Develop an Open-source Multimodal Tool

The third and final goal of this work is to develop and release an open-source tool. This thesis presents the **OpenMM** tool: a multimodal feature extraction tool. To the best of our knowledge, this tool is the first of its kind. We hope this tool will allow researchers to easily extract features from various modalities all at once. History has shown that major factors that help facilitate collaboration are the sharing and making public of both datasets and code. In the past, some researchers have made their datasets public, including corpora from the Audio/Visual Emotion Challenges (Valstar et al., 2013, 2014, 2016a). Due to these publicly-released corpora, a good deal of impactful research has been conducted and interest in depression detection has grown. In addition to publicly-releasing corpora, sharing code is a useful practice to help increase collaboration, cooperation, and reproducible research. Many researchers have created useful open-source tools that have benefited the community, including Covarep (Degottex et al., 2014) and OpenFace (Amos et al., 2016). OpenMM leverages existing tools as well as newly developed code to present a multimodal feature extraction pipeline. By providing a simple and efficient way to easily extract multimodal features, we hope to see an increase of interest in multimodal systems, which in turn can promote collaboration between the fields.

1.2 Included Works

This thesis is a compendium of published works (Morales and Levitan, 2016a,b; Morales et al., 2017a,b). Each included work addresses one or more of the research goals outlined above. Each work receives its own chapter and the following sub-sections provide a short description of each. Each chapter dedicated to an included published work (Chapter 3, 4, 5, and 6) includes a relevant related work, methodology, results, and discussion section. In addition to the included work chapters, we also present a theoretical background in Chapter 2 and novel unpublished results in Chapter 7. We conclude this thesis with a thorough discussion of limitations, contributions, and future work, in Chapter 8.

1.2.1 A Cross-modal Review of Approaches for Depression Detection Systems

In Chapter 3, we present a survey of depression detection systems across fields. This survey is the first to present a thorough cross-modal review of depression detection systems. The review discusses best practices and most promising approaches to this task.

1.2.2 A Comparative Analysis of Features for Depression Detection Systems

In Chapter 4, we provide a comparative analyses of various features for depression detection. Using the same corpus, we evaluate how a system built on text-based features compares to a speech-based system. We find that a combination of features drawn from both speech and text lead to the best system performance.

1.2.3 An Open-source Multimodal Feature Extraction Tool

In Chapter 5, we present OpenMM: an open-source multimodal feature extraction tool. We build upon existing open-source repositories to present the first publicly available tool for multimodal feature extraction. The tool provides a pipeline for researchers to easily extract visual and acoustic features. The tool also performs automatic speech recognition (ASR) and then uses the transcripts to extract linguistic features. We evaluate the OpenMM's multimodal feature set on depression detection. In order to demonstrate the tool's robustness, we also evaluate it on other related machine learning tasks, including deception and sentiment detection. Across all tasks we show OpenMM's performance is very promising.

1.2.4 Mitigating Confounding Factors in Depression Detection Using an Unsupervised Clustering Approach

In Chapter 6, we discuss challenges to depression detection, specifically the presence of confounding factors, such as gender, age, emotion and personality. We discuss approaches to handling confounding factors and present a technique to mitigate such factors, which uses a multi-step approach that performs unsupervised clustering prior to depression classification.

1.3 Contributions

In the pursuit of addressing our research goals, this thesis contributes the following to the depression detection research community:

1. Theoretical background on depression including relevant psychology and linguistic literature (Chapter 2)
2. An empirical survey of depression detection systems across research fields and modalities (Chapter 3)
3. A systematic evaluation of unimodal and multimodal depression detection systems (Chapters 4, 5, 6, and 7)
4. A presentation and evaluation of a novel multimodal feature extraction tool (Chapter 5)
5. A thorough investigation of fusion techniques for multimodal depression detection systems (Chapter 7)
6. A presentation and evaluation of a novel fusion technique (Chapter 7)

Chapter 2

Background

This chapter discusses relevant literature from psychology and linguistics. In Section 2.1, we present the psychological background, which covers the clinical definition of depression, diagnostic methods, treatments, objective markers, and theories of depression. In Section 2.2, we present a linguistic background, which includes an overview of the language production process as well as the speech production system. In addition, we cover relevant literature on depression and its influence on linguistic behavior.

2.1 Psychological Background

2.1.1 Clinical Definition of Depression

Clinical depression is a psychiatric mood disorder that is caused by an individual's inability to cope with certain stressful events and situations. According to the Diagnostic and Statistical Manual of Mental Disorders (DSM-5), the most widely used resource in diagnosing mental disorders in the United States, most people will feel some form of depression in their lifetime, although it does not reach the level of an illness until a person has experienced, for longer than a two-week period, a depressed mood and/or a markedly diminished interest/pleasure in combination with four or more of the symptoms listed in Table 2.1 (American Psychological Association [APA], 2013).

Table 2.1: DSM-5 criteria for depression.

Symptoms
Depressed mood and/or loss of interest in pleasure, in combination with 4 or more of the following:
Significant unintentional weight loss or gain
Insomnia or sleeping too much
Agitation or psychomotor retardation noticed by others
Fatigue or loss of energy
Feelings of worthlessness or excessive guilt
Diminished ability to think or concentrate, or indecisiveness
Recurrent thoughts of death

The current criteria for major depression, given in Table 2.1, have been criticized for the heterogeneity of the clinical syndrome it defines. According to these criteria, there are actually at least 1,497 unique profiles of depression. Therefore, two individuals can present with completely different symptoms and receive the same diagnosis. Due to this heterogeneity, some have suggested that a redefinition of the depressive syndrome may be warranted (Østergaard et al., 2011). However, given that these criteria constitute the most standard definition in the United States, we adopt them here.

2.1.2 Diagnostic Methods and Treatments for Depression

The diagnosis of depression in primary care settings is difficult and findings have shown that physicians both commonly under and over diagnose depression (Schumann et al., 2012). Factors that complicate diagnosis include the time consuming nature of diagnosis and the modest rate of clinical depression seen in primary care settings. Moreover, not all depressed patients outwardly express emotional symptoms. Commonly used assessment tools for depression include clinical interviews, rating scales, and self-assessments. The Hamilton Rating Scale for Depression (HAMD; Hamilton, 1960), which has been shown to have predictive validity and consistency, is a widely used assessment tool and is often regarded as the most standard assessment tool for depression for both diagnosis and research purposes (Cummins et al., 2015a). The HAMD is clinician-administered, includes 21 questions, and takes 20 to 30 minutes to complete. The interview assesses the severity of symptoms associated with depression and gives a patient a score, which relates to their level of depression. Symptoms include but are not limited to depressed mood, insomnia, agitation, and anxiety. Each of the questions has 3 to 5 possible responses which range in severity, scored between 0-2, 2-3, or 4-5 depending on the importance of the symptom. All scores are then summed and the total is arranged into 5 categories (from normal to severe).

On the other hand, the most widely used self-reported measure is the Beck Depression Inventory (BDI; Beck et al., 1961). The BDI consists of 21 items and takes 5 to 10 minutes to complete. The question items aim to cover important cognitive, affective, and somatic symptoms observed in depression. Each question receives a score on a scale from 0-3 depending on how severe the symptom was over the previous week. Similar to HAMD, all scores are summed and the final score is categorized into 4 different levels (from minimal to severe). Nuevo et al. (2009) showed that the BDI is reliable and valid for use in the general population. Other diagnostic tools include: Montgomery-Åsberg Depression Rating Scale

(Montgomery and Asberg, 1979), Patient Health Questionnaire (Kroenke et al., 2001), and Quick Inventory of Depressive Symptomology (Rush et al., 2003).

Depression is a treatable illness. Depending on the type of depression, various treatments are available, including medication and several types of therapy.

2.1.3 Objective Markers for Depression

Due to the variation in depression profiles, it is evident that the diagnosis of depression is difficult. Therefore, researchers have extensively investigated possible objective markers of depression. Although objective markers have not been fully accepted as diagnostic measures, they have a vast range of potential uses in psychology. Markers include biological, physiological, and behavioral markers.

In regards to biological markers for depression, long-standing theory suggests that a breakdown in brain serotonin signaling is critically involved in the symptoms and treatment of clinical depression (Sharp and Cowen, 2011). Although lower levels of serotonin are reported for those with depression, it is not a specific marker for the depressed and can appear in healthy individuals. However, a lower level of serotonin can arguably represent a mental state vulnerability and thus can be associated with depression. In addition, low functioning of the neurotransmitter GABA (gamma-amino butyric acid) has also been linked with a vulnerability to depression (Croarkin et al., 2011). Other biological markers for depression include molecular markers, such as insulin and serum (Schmidt et al., 2011), protein molecules such as cytokines (Schmidt et al., 2011), and steroid hormones such as salivary cortisol levels (Owens et al., 2014). In addition to biological markers for depression, researchers have also investigated physiological markers. Studies have found physiological markers of depression, including galvanic skin responses (Schneider et al., 2012), saccadic eye movements (Steiger and Kimura, 2010), changes in REM sleep parameters (Hasler et al., 2004), and cardiovascular dysregulation (Carney et al., 2005).

Table 2.2: Behavioral markers associated with depression.

Behavioral Signal	Effect	References
Social Interaction	↓	Bos et al. (2002) and Hall and Rosenthal (2005)
Clinical Interaction	↓	Parker et al. (1990)
Gross Motor Activity	↓	Balsters and Vingerhoets (2012), Parker et al. (1990), and Sobin and Sackeim (1997)
Slumped Posture	↑	Parker et al. (1990) and Segrin (2000)
Gesturing	↓	Balsters and Vingerhoets (2012) and Segrin (2000)
Self-touching	↑	Scherer et al. (2013b), Segrin (2000) and Sobin and Sackeim (1997)
Head-movement Variability	↓	Girard et al. (2014) and Scherer et al. (2013c)
Mobility	↓	Parker et al. (1990) and Sobin and Sackeim (1997)
Expressivity	↓	Ellgring and Scherer (1996), Gaebel and Wolwer (2004), Girard et al. (2014), Maddage et al. (2009), Schelde (1998), and Segrin (2000)
Smiling	↓	Balsters and Vingerhoets (2012), Schelde (1998), Scherer et al. (2013b), Segrin (2000), and Sobin and Sackeim (1997)
Eyebrow Movements	↓	Balsters and Vingerhoets (2012), Schelde (1998), and Segrin (2000)
Visual Fixation	↑	Abel et al. (1991) and Lipton et al. (1980)
Saccades	↓	Sweeney et al. (1998)

Note. Table taken from Cummins et al. (2015a). ↓ indicates a reduction in the behavior while ↑ indicates an increase in the behavior.

In addition to biological and physiological markers, a good deal of psychological research has investigated behavioral markers associated with clinical depression. Table 2.2, borrowed from Cummins et al.'s (2015a) extensive review of depression assessment research, outlines how behaviors are affected by depression, giving insight into how depression manifests within individuals. For example, depression can be associated with a decrease of social interaction and gesturing, while also associated with an increase of slumped posture and self-touching.

2.1.4 Theories of Depression

In this sub-section we discuss psychological theories of depression including psychoanalytic, cognitive, behaviorist, and self-aware theories. Since Freud's seminal work (Freud, 1917), psychoanalytic theories of depression have asserted the importance of loss in the onset of depression, specifically loss of love and emotional security. Freud observed that losses produce severe and irrational self-criticism of oneself. Freud's psychoanalytic theory is one of the many existing perspectives on depression. In later work, Freud modified his theory stating that the tendency to internalize loss is normal. He then posited that depression is simply due to an excessively severe super-ego. Depression occurs when the individual's super-ego or conscience is dominant.

In contrast to Freud's theories, Aaron Beck's (1967) cognitive theory of depression posits that three mechanisms are responsible for depression (McLeod, 2015):

1. The cognitive triad of negative automatic thinking
2. Negative self schemas
3. Errors in logic (i.e. faulty information processing)

The first mechanism, the cognitive triad, represents three forms of negative thinking that are typical of individuals with depression. The triad is made up of negative thoughts about

(1) the self, (2) the world and (3) the future. Beck argued that these thoughts tend to be automatic in depressed people, occurring spontaneously. For example, depressed individuals tend to view themselves as helpless, worthless, and inadequate. They also tend to interpret events in the world in a unrealistically negative way and finally, they see the future as totally hopeless.

The second mechanism Beck posited was that people prone to depression possess a depressive schema, or a deep level knowledge structure, that leads them to see themselves and the world in pervasively negative terms. These schemas when activated give rise to depressive thinking. Consequently, stressful events can trigger depressive schemas leading an individual to perceive an event in a negative way, causing an episode of depression. These negative schemas seem to then lead to the third and final mechanism: errors in logic. Beck argued that once the negative schema are activated a number of illogical thoughts or cognitive biases seem to dominate thinking. As a result, people with negative self schemas become prone to making logical errors in their thinking, focusing selectively on certain aspects of a situation while ignoring equally relevant information (McLeod, 2015).

Behaviorist theories of depression provide an additional perspective, which emphasizes the importance of environment. Behaviorist theory focuses on the observable behavior and the conditions through which individuals learn behavior. Behaviors can be learned in many ways, such as classical conditioning, operant conditioning and social learning theory. Under behaviorist theory, depression is the result of a person's interaction with their environment. For example, under classical conditioning, depression is learned through associating certain stimuli with negative emotional states. Under social learning theory, depression is learned through observation, imitation, and reinforcement. Under operant conditioning, depression is caused by the removal of positive reinforcement from the environment. For example, Lewinsohn (1974) argued that certain events, such as losing your job, induce depression because they reduce positive reinforcement from others.

Pyszczynski and Greenberg (1987) proposed the self-awareness theory for depression, which speculated that depressed individuals think a great deal about themselves, stressing the role of self-focused attention. After the loss of a central source of self-worth, individuals are unable to exit a self-regulatory cycle concerned with efforts to regain what was lost. This results in self-focus, which in turn is thought to magnify negative emotion and self-blame. Therefore, the depressive self-focusing style maintains and exacerbates depression. Tangentially related work by Durkheim (1951) posits a social integration theory of suicide, in which the perception of oneself as not integrated into society (detached from social life) is key to suicidality and also relevant to the depressed person's perceptions of self.

In summary, each theory presents a different perspective on the cause and manifestation of depression. However, the common theme seen throughout many theories is the notion that depression influences how a person perceives themselves and the world. Undoubtedly, this perception can have dramatic effects on an person's behavior and language.

2.2 Linguistic Background

Given existing psychological theories of depression, both psychologists and linguists have investigated how these theories could manifest in language. We know language is a medium; it is the most common and reliable way for people to translate their internal thoughts and feelings into a form that others can understand (Tausczik and Pennebaker, 2010). Therefore, language is the medium by which cognitive, personality, clinical, and social psychologists attempt to understand human beings (Tausczik and Pennebaker, 2010). Undoubtedly, depression affects and influences the way individuals feel, think, and communicate. Therefore, by analyzing a person's language systematically, we can learn how depression influences their feelings and state of mind. However, understanding linguistic behavior and patterns is difficult. If we consider the process of information flow during language production, we note that

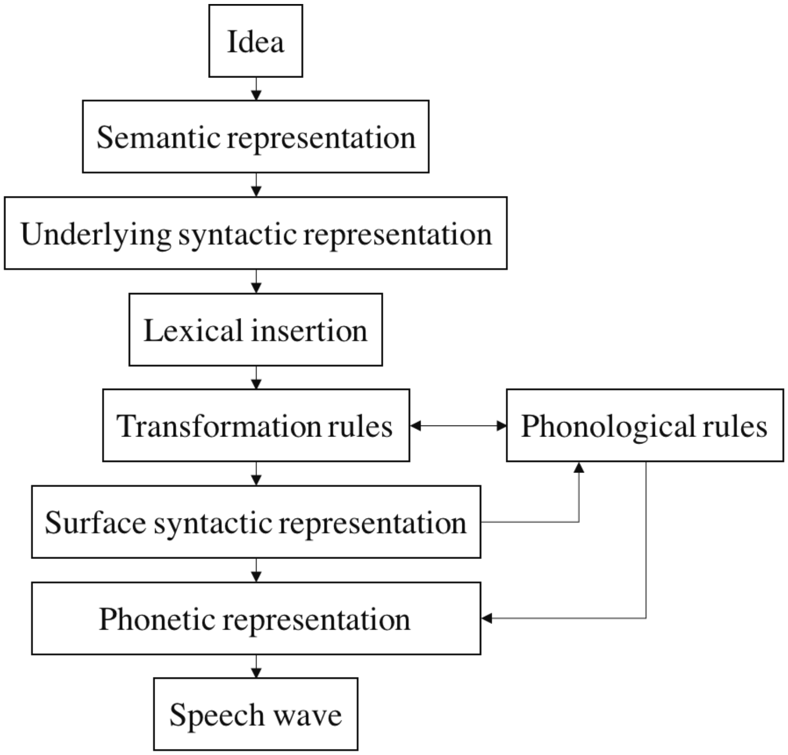


Figure 2.1: Model of information flow during speech production, taken from Cooper and Paccia-Cooper's work on speech and syntax (Cooper and Paccia-Cooper, 1980).

the entire set of operations by which a speaker transforms ideas into acoustic output is enormously complex (Cooper and Paccia-Cooper, 1980; Lenneberg et al., 1967). For example, see Figure 2.1, which represents a model of information flow during speech production[1].

According to this model, when a speaker plans to produce a meaningful utterance, information is processed at a number of different levels. At the first stage, a speaker generates an idea, which is then translated into a linguistic form or semantic representation. The speaker then formulates a partial grammatical representation of the utterance. This is followed by the speaker choosing one or more major lexical items. When a fully elaborated underlying structure has been formulated, it is assumed that the structure may undergo transformations that move/add/delete constituents. Output from the transformation stage comprises

[1] Important to note, this model represents a simplified schematization of the language production process and is just meant to serve as an overview of the process.

the surface structure. Then, phonological rules of stress assignment may apply to the output of the surface structure. Finally, phonetic representations are transferred to the motor system, which generates the articulatory configurations of speech, i.e. the acoustic output. Considering the complexity of the information flow during the speech production process, it is clear that many stages of the process can ultimately affect a speaker's acoustic output. We know that the stages of semantics and syntax affect speech (Cooper and Paccia-Cooper, 1980). We also know that a multitude of psychological symptoms can affect language behavior (Blanken et al., 1993). Given this tiered process of production, it is necessary to analyze language at different levels.

One approach to analyzing a person's language is via text analysis studies. Text analysis dates back to the earliest days of psychology, where Freud wrote about slips of the tongue (Freudian slips) and how these apparent linguistic mistakes reveal the true secret thoughts and feelings that people hold (Freud, 1901). Text analysis studies related to depression have uncovered interesting relationships between depression and language use. For example, Stirman and Pennebaker (2001) studied the word usage of suicidal and non-suicidal poets. They conducted a comparison of 300 poems from the early, middle, and late periods of nine poets who committed suicide and nine who did not. Using the Linguistic Inquiry and Word Count dictionary (LIWC; Pennebaker et al., 2007), they found suicidal poets used more first-person singular (*I, me, my*) words, and fewer words pertaining to the social collective (*we, us, our*). Interestingly, the two groups did not differ in negative or positive emotion words. Important to note, it is unclear whether their findings were due to the poets' suicidality or depression, or both, but other studies have noted the elevated use of first person singular pronouns by depressed persons (Bucci and Freedman, 1981). The increased use of first person singular pronouns suggests a tendency of depressed individuals to focus mostly on themselves. Their findings provided evidence consistent with both the self-awareness and social integration theories.

In later work, Rude et al. (2004) analyzed narratives written by currently-depressed, formerly-depressed, and never-depressed college students. Their work provided an interesting insight into how depression-prone individuals (those with a history of depression) think at times when they are not experiencing an episode. They examined linguistic patterns of depressed and depression prone students in the context of an essay task. The prompt stated: *write about your deepest thoughts and feelings about coming to college.* The linguistic characteristics they analyzed were the following:

1. first person singular (*I, me, my*)

2. first person plural (*we, us, our*)

3. social references (*mention of friends, family, relationship titles, etc.*)

4. negatively valenced words (*gloom, flight, sad, homesick, etc.*)

5. positively valenced words (*joyful, accept, best, play, share, etc.*)

As predicted by Pyszczynski and Greenberg's (1987) self-awareness theory, depressed students used significantly more first person singular pronouns than did never-depressed individuals. They also found that depressed students used more negatively valenced words and fewer positive emotion words. Therefore, depressed students revealed both the negative focus predicted by Beck's cognitive theory of depression and the self-preoccupation predicted by Pyszczynski and Greenberg's theory of depression. However, there was no evidence of social isolation/disengagement as would be predicted by Durkheim's model of suicidality in social references among the depressed students; this could have been due to the writing topic, which biased students against writing their experiences as part of a group.

In addition to focusing on lexical (word) patterns, text analysis studies have also investigated depression and syntax (Zinken et al., 2010). Zinken et al. investigated whether an analysis of a depressed patient's syntax could help predict improvement of depressed symptoms. They built upon previous work that found health benefit from 'expressive' writing; health benefit was positively correlated with an increase in the use of causation words (such as *because* or *effect*) and insight words (such as *think* or *consider*) over writing sessions (Pennebaker, 1997). Zinken et al. considered the psychological relevance of syntactic structures of language use, noting that texts can barely differ in their word usage, but they may differ in their syntactic structure, and consequently in the construction of relationships between events. Word use and syntactic structure were analyzed to explore whether the degrees to which a participant constructs relationships between events in a brief text could inform the likelihood of successful participation in guided self-help. They hypothesized that the use of causation and insight words and of complex syntactic structure would help to predict (1) completion of the program and (2) benefit of the program. Using LIWC, they targeted two categories: causation words and insight words. In addition, they manually coded eight different syntactic structures (representing an exhaustive set of the grammatically possible cross-clausal relationships) in the patients' narratives. Zinken et al. found that certain structures were correlated with patients' potential to complete a self-help program and benefit from it. The use of complex syntax, i.e. adverbial clause use, predicted improvement. Therefore, those individuals that remained depressed were less likely to employ complex syntactic constructions.

In addition, linguists have long postulated the important relationship between syntax and prosody (Chomsky and Halle, 1968). Some have even argued that prosody can be directly predicted from the syntactic tree configuration of a sentence (Wagner, 2004). Therefore, it is important to not only consider word use and syntax, but also prosodic and more general phonetic characteristics of language use. The speech production system of a human is very complex. Lenneberg et al. (1967) estimated that more than 100 independently innervated muscles are coordinated in the tongue and mouth during speech. As a result of the system's

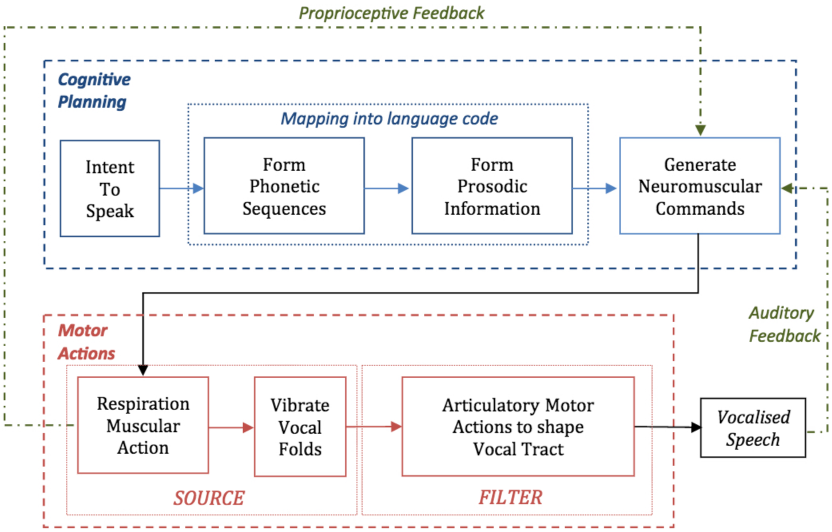

Proprioceptive Feedback
Cognitive Planning
Mapping into language code
Intent To Speak
Form Phonetic Sequences
Form Prosodic Information
Generate Neuromuscular Commands
Auditory Feedback
Motor Actions
Respiration Muscular Action
Vibrate Vocal Folds
Articulatory Motor Actions to shape Vocal Tract
Vocalised Speech
SOURCE
FILTER

Figure 2.2: Schematic diagram of speech production taken from Cummins et al. (2015a).

complexity, speech is sensitive. Some have argued that slight physiological and cognitive changes can produce acoustic changes in speech (Scherer, 1986). In this work, we assume that depression produces cognitive and physiological changes that influence speech production, leading to a change in the acoustic quality of the speech produced. This change can then be measured and objectively evaluated. Understanding the speech production system, can help shed light on how certain components can be affected by the presence of depression.

When we wish to communicate, we first cognitively plan a message. We then establish the phonetic and prosodic information of the message, which is stored in short-term working memory. This information is then transformed into the actual phonetic and prosodic representations and the speaker can then execute a series of neuromuscular commands. These commands initiate the motoric actions needed to produce speech. Motor actions are made up of the *source* and *filter*. The source is air produced by the lungs, which passes through the filter which shapes the sound. The filter, given in Figure 2.2, represents the vocal tract. The articulators of the vocal tract, such as the glottis, alter the sound or phoneme produced by how they are positioned. Studies have investigated the cognitive effects on speech production. Subsequently, some have found that cognitive impairments associated with depression have an effect on an individual's working memory. The phonological loop, shown in Figure 2.2, is a key component in the speech production system and is part of working memory; the loop is responsible for helping control the articulatory system. Therefore, a cognitive impairment on working memory can affect this part of the speech production system. Psychological research has confirmed this; Christopher and MacDonald (2005) showed that depression affects the phonological loop causing articulation and phonation errors. Consequently, the effects of depression reflected in the speech production system makes speech an attractive candidate for an objective marker of depression.

2.3 Summary

According to its clinical definition, depression is an extremely heterogeneous disorder which is difficult and time-consuming to diagnose. Therefore, technology, namely detection systems, which can automatically monitor and analyze depression are extremely attractive because of their efficiency, systematicity, and objectiveness. The detection systems presented in this work are motivated from the included literature. Psychologists and linguists have shown that depression influences how a person behaves and communicates. Both nonverbal and verbal behavior have been shown to be affected by depression, including facial expressions (Cummins et al., 2015a), prosody (Hönig et al., 2014; Mundt et al., 2012; Trevino et al., 2011; Yang et al., 2013), syntax (Zinken et al., 2010), and semantics (Oxman et al., 1988; Rude et al., 2004). Therefore, these theories and studies motivate our multimodal system: a depression detection system that captures as many of these characteristics and patterns as feasible. A multimodal detection system provides the ideal framework to capture the many ways in which depression may influence a person. In the next chapter, we present our comprehensive cross-modal empirical review of depression detection systems.

Chapter 3

A Cross-modal Review of Approaches for Depression Detection Systems

3.1 Motivation

This chapter presents published work[1]. As outlined in Chapter 1, one major challenge facing the depression detection research community is the existing disconnect between subfields: approaches to depression assessment from NLP, speech processing, and HCI tend to silo by subfield, with little discussion about the utility of combining promising approaches. This existing disconnect necessitates a bridge to facilitate greater collaboration and cooperation across subfields and modalities. This work aims to serve as a bridge between the subfields by providing the first review of depression detection systems across modalities. In this chapter, we focus on the following research questions: how has depression been defined and annotated in detection systems? What kinds of depression data exist or could be obtained for depression detection systems? What (multimodal) indicators have been used for the

[1] Morales, M. R., Scherer, S., and Levitan, R. (2017a). A Cross-modal Review of Indicators for Depression Detection Systems. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology*, pages 1-12, Vancouver, Canada. Association for Computational Linguistics.

automatic detection of depression? How do we evaluate depression detection systems? In addition, we discuss factors that require attention when building systems for depression detection.

3.2 Defining and Labeling Depression

3.2.1 Clinical Definition and Diagnostics

As described in Chapter 2, most people will experience some feelings of depression in their lifetime, although it does not meet the criteria of an illness until a person has experienced, for longer than a two-week period, a depressed mood and/or a markedly diminished interest/pleasure in combination with four or more of the following symptoms: significant unintentional weight loss or gain, insomnia or sleeping too much, agitation or psychomotor retardation noticed by others, fatigue or loss of energy, feelings of worthlessness or excessive guilt, diminished ability to think or concentrate, indecisiveness, or recurrent thoughts of death (APA, 2013). Diagnosis requires that the symptoms cause clinically significant distress or impairment in social, occupational, or other important areas of functioning. In addition, commonly used assessment tools for depression include clinical interviews, rating scales, or self-assessments.

3.2.2 Scalable Approaches to Annotation

When working with datasets, it is not always feasible to acquire clinical ratings for depression level. As a result, researchers have come up with innovative ways of acquiring depression labels at scale, notably from social media sources. Given the explosion of social media, this domain is especially rich in data for mental health research. However, any research in this domain must take into account the ability of online users to be anonymous or even deceptive.

Coppersmith et al. (2015) looked for tweets that explicitly stated “I was just diagnosed with depression”. Moreno et al. (2011) evaluated Facebook status updates using references to depression symptoms such as “I feel hopeless” to ultimately determine depression label. Choudhury et al. (2013) used crowdsourcing, via the Amazon Mechanical Turk platform, to collect Twitter usernames as well as labels for depression. Reece and Danforth (2016) used a similar crowdsourcing approach to collect both depression labels and Instagram photo data. In some approaches to annotation, depression is subsumed into broader categories like distress, anxiety, or crisis. For example, Milne et al. (2016) used judges to manually annotate how urgently a blog post required attention, using a triage system of green/amber/red/crisis.

These innovative approaches to data annotation highlight the potential of social media data. This domain offers a very rich data source which can be used to build, train, and test models to automatically perform mental health assessments at scale.

3.3 Datasets

The task of depression detection is inherently interdisciplinary and all disciplines—psychology, computer science, linguistics—bring an essential set of skills and insight to the problem. However, it is not always the case that a team is fortunate enough to have collaborators from all disciplines. One way to promote collaboration is to organize challenges and publicly release data and code. Public datasets are invaluable resources that can give new researchers the ability to work on the task while connecting accomplished researchers across disciplines. The Computational Linguistics and Clinical Psychology (CLPsych) Shared Task (2013-2017) and the Audio/Visual Emotion Recognition (AVEC) Depression Sub-challenge (2013-2016) are examples of depression detection system challenges that spurred interest, promoted research, and built connections across the research community. In this section, we describe the kinds of depression data that exist, listed in Table 3.1. We focus solely on datasets that are publicly available to download. For a detailed list of databases both private and public that have been used in speech processing studies see Cummins et al. (2015a).

Table 3.1: Datasets for depression detection systems

Dataset	Modality	Depression Annotation	Reference
AVEC 2013	A + V	BDI-II	(Valstar et al., 2013)
AVEC 2014	A + V	BDI-II	(Valstar et al., 2014)
Crisis Text Line	T	Manual annotation for depression	(Lieberman and Meyer, 2014)
DAIC	A + T + V	PHQ-8	(Gratch et al., 2014)
DementiaBank Database	A + T + V	HAMD	(Becker et al., 1994)
ReachOut Triage Shared Task	T	Manual annotation for triage label	(Milne et al., 2016)
SemEval-2014 Task 7	T	Manual annotation for depression	(Pradhan et al., 2014)

Note. A represents audio, T represents text, and V represents video.

3.3.1 AVEC 2013/2014

Both the AVEC 2013 and 2014 corpora are available to download[2]. The AVEC challenges are organized competitions aimed at comparing multimedia processing and machine learning methods for automatic audio, video and audiovisual emotion and depression analysis, with all

[2] https://avec2013-db.sspnet.eu/

participants competing strictly under the same conditions. The AVEC 2013 corpus includes 340 video clips in German of subjects performing a HCI task while being recorded by a webcam and a microphone (Valstar et al., 2013). The video files each contain a range of vocal exercises, including free and read speech tasks. The level of depression is labeled with a single value per recording using the BDI-II. The AVEC 2014 corpus (Valstar et al., 2014) is a subset of the AVEC 2013 corpus. In total, the corpus includes 300 videos in German, with duration ranging from 6 seconds to 4 minutes. The files include a read speech passage (Die Sonne und der Wind) and an answer to a free response question.

3.3.2 Crisis Text Line

The Crisis Text Line [3] is a free 24/7 crisis support texting hot line where live trained crisis counselors receive and respond quickly to texts. The main goal of the organization is to support people with mental health issues through texting. The organization includes an open data collaboration. In order to gain access, researchers must complete an Institutional Review Board application with their own university and an application with Crisis Text Line, which gives researchers access to a vast amount of text data annotated by conversation issue, including but not limited to depression, anger, sadness, body image, homelessness, self-harm, suicidal ideation, and more.

3.3.3 DAIC

The Distress Analysis Interview Corpus (DAIC; Gratch et al., 2014) contains clinical interviews in English designed to support the diagnosis of psychological distress conditions such as anxiety, depression, and post-traumatic stress disorder. The interviews were conducted by an animated virtual interviewer called Ellie. The DAIC interviews were meant to

[3] www.crisistextline.org

simulate the first step in identifying mental illness in health care settings, which is a semi-structured interview where health care providers ask a series of open-ended questions with the intent of identifying clinical symptoms. The corpus includes audio and video recordings and extensive questionnaire responses. Each interview includes a depression score from the Patient Health Questionnaire-8 (PHQ-8; Kroenke et al., 2009). A portion of the corpus was released during the AVEC 2016 Depression Sub-challenge and is available to download[4]. The publicly-available dataset also includes transcripts of the interview.

3.3.4 DementiaBank

The DementiaBank Database[5] represents data collected between 1983 and 1988 as part of the Alzheimer Research Program at the University of Pittsburgh (Becker et al., 1994). DementiaBank is a shared database of multimedia interactions for the study of communication in dementia. A subset of the participants from the dataset also have HAMD depression scores.

3.3.5 ReachOut Triage Shared Task Dataset

The ReachOut Triage Shared Task dataset[6] consists of 65,024 forum posts written between July 2012 and June 2015 (Milne et al., 2016). A subset of the corpus (1,227 posts) is manually annotated by three separate expert judges indicating how urgently a post required a moderator's attention. These triage labels include crisis, red, amber, and green.

[4] http://dcapswoz.ict.usc.edu/

[5] http://dementia.talkbank.org/

[6] http://clpsych.org/shared-task-2016/

3.3.6 SemEval-2014 Task 7

The SemEval-2014 Task 7 (Pradhan et al., 2014) dataset[7] represents clinical notes which are annotated for disorder mentions, including mental disorders such as depression.

3.4 Indicators of Depression

Ideally, machine learning tools for depression detection should have access to the same streams of information that a clinician utilizes in the process of forming a diagnosis. Therefore, features used by such classifiers should represent each communicative modality: face and gesture, voice and speech, and language. This section provides a review of each modality highlighting markers that have had success in depression detection systems.

3.4.1 Visual Indicators

Visual indicators have been widely explored for depression analysis, including body movements, gestures, subtle expressions, and periodical muscular movements.

Girard et al. (2014) investigated whether a relationship existed between nonverbal behavior and depression severity. In order to measure nonverbal behavior they used the Facial Action Coding System (FACS; Ekman et al., 1978). FACS is a system used to taxonomize human facial movements by their appearance on the face. It is a commonly used tool and has become standard to systematically categorize physical expressions, which has proven very useful for psychologists. FACS is composed of facial Action Units (AUs), which represent the fundamental actions of individual muscles or groups of muscles. Girard et al. (2014) found that participants with high levels of depression made fewer affiliative facial expressions, more non-affiliative facial expressions, and diminished head motions. Scherer et al. (2013b) also investigated visual features using FACS and found that depression could be predicted by

[7] http://alt.qcri.org/semeval2014/task7/index.php?id=data-and-tools

a more downward angle of the gaze, less intense smiles, shorter average durations of smile, longer self-touches, and fidgeting.

In addition to FACS features for video analysis, others have considered Space-Time Interest Points (STIP) features (Cummins et al., 2013; Joshi et al., 2013b), which capture spatio-temporal changes including movements of the face, hands, shoulder, and head. Using STIP features, Joshi et al. (2013b) found that they could detect depression with 76.7% accuracy. Their results showed that body expressions, gestures, and head movements can be significant visual cues for depression detection.

3.4.2 Speech Indicators

In patients with depression, several changes in prosody have been noted, including reduced dynamics of loudness, narrower pitch range, and reduced variation of pitch (standard deviation of pitch) (Blanken et al., 1993). Early studies into depressed speech found patients consistently demonstrated prosodic abnormalities, such as reduced pitch, slower speaking rate, and articulation errors. Darby and Hollien (1977) found that listeners noted differences in pitch, loudness, speaking rate, and articulation of depressed individuals before and after treatment. They had 112 listeners evaluate voice recordings containing fundamental frequency contours of individuals during depression and after treatment. In 80% of the cases, listeners were able to identify whether or not the recording was from a phase of depression. Many researchers have noted correlations between a reduced F_0 range and a reduced F_0 average with increasing levels of depression. However, a number of studies report contrary results, showing no significant correlations between F_0 variables and depression level. These conflicting results could be attributed to the heterogeneous nature of how depression presents itself in individuals.

Recent research has shown the promise in using speech as a diagnostic and monitoring aid for depression (Cummins et al., 2015b,a, 2014; Scherer et al., 2014; Williamson et al., 2014a). The speech production system of a human is very complex and as a result slight cognitive or physiological changes can produce acoustic changes in speech. This idea has driven the research on using speech as an objective marker for depression. Depressed speech has consistently been associated with a wide range of prosodic, source, formant and spectral indicators. For a thorough review of speech processing research for depression detection see Cummins et al. (2015a).

Many researchers have provided evidence for the robustness of prosodic indicators to capture depression level, specifically noting the promise of speech-rate (Hönig et al., 2014; Mundt et al., 2012). Cannizzaro et al. (2004) examined the relationship between depression and speech by performing statistical analyses of different acoustic measures, including speaking rate, percent pause time, and pitch variation. Their results demonstrated that speaking rate and pitch variation had a strong correlation with the depression rating scale. Moore et al. (2008) investigated the suitability for a classification system formed from the combination of prosodic, voice quality, spectral, and glottal features and reported maximum accuracy of 91% for male speakers and 96% accuracy for females speakers when classifying between absence/presence of depression.

Stassen et al. (1998) found for 60% of patients in their study that speech pause duration was significantly correlated with their HAMD score. Alpert et al. (2001) also found significant differences in speech pause duration between spontaneous speech of their depressed group versus their control group. Cannizzaro et al. (2004) found a significant correlation between reduced speaking rate and HAMD score. Mundt et al. (2012) found six prosodic timing measures to be significantly correlated with depression severity, including total speech time, total pause time, percentage pause time, speech pause ratio, and speaking rate. Hönig et al. (2014) reported a positive correlation with increasing levels of speaker depression and average syllable duration. Trevino et al. (2011) found that changes in speech rate are stronger at the phoneme level, finding stronger relationships between speech rate and depression severity when using phone-duration and phone-specific measures instead of a global speech rate. Cohn et al. (2009) investigated vocal prosody and found that variation in fundamental frequency and latency of response to interviewer questions achieved 79% accuracy in distinguishing participants with moderate/severe depression from those with no depression.

Low et al. (2011) investigated various acoustic features, including spectral, cepstral, prosodic, glottal and a Teager energy operator based feature. In their best performing systems, using sex-dependent models, they achieved 87% accuracy for males and 79% for females. In Cummins et al. (2011) spectral features, particularly mel-frequency cepstral coefficients (MFCCs) were found to be useful, distinguishing 23 depressed participants from 24 controls with an accuracy of 80% in a speaker-dependent configuration. Scherer et al. (2013a) found glottal features (normalized amplitude quotient and quasi-open quotient) differed significantly between depressed and control groups. When used to detect depression they found glottal features to differentiate between the 2 groups with 75% accuracy. Alghowinem et al. (2013) investigated a number of feature sets for detecting depression from spontaneous speech and found loudness and intensity features to be the most discriminative.

3.4.3 Linguistic and Social Indicators

While most literature concerning depression detection systems has focused on the speech signal, there is a related body of work on detecting depression from writing using linguistic cues. For clinical psychologists, language plays a central role in diagnosis. Therefore, when building language technology in the domain of mental health it is essential to consider both the acoustic and linguistic signal. For an in-depth review of NLP applications for mental health assessment see Calvo et al. (2017a).

Features derived from the speech signal are motivated by ways in which the cognitive and physical changes associated with depression can lead to differences in speech. Similarly, psychological and sociological theories suggest that depressed language can be characterized by specific linguistic features. As discussed in Chapter 2, Section 2.1.4, many theories of depression exist, including Aaron Beck's (1967) cognitive theory of depression, Pyszczynski and Greenberg's (1987) self-awareness theory, and Durkheim's (1951) social integration model. These theories have motivated empirical studies of depressed language which have in turn provided support for their validity. As described in Chapter 2, Section 2.2, many researchers have investigated patterns in word choice using the tool LIWC. LIWC is a text analysis tool that can be used to count words in psychologically meaningful categories (Tausczik and Pennebaker, 2010). Early text analysis research found that depressed individuals used significantly more first person singular words than did never-depressed individuals as well as more negatively valenced words than positive emotion words (Pennebaker et al., 2007; Rude et al., 2004; Stirman and Pennebaker, 2001). Given the success of LIWC in text analysis studies, many other researchers have incorporated LIWC into depression detection systems with encouraging results. Nguyen et al. (2014) used LIWC to capture topic and mood of depressed individuals' writing. They found LIWC features showed good predictive validity in depression classification between clinical and control groups in blog post texts. Morales and Levitan (2016b) incorporated LIWC into a depression detection system and found certain LIWC categories to be useful in measuring specific depression symptoms, including sadness and fatigue.

In addition to LIWC, researchers have also experimented with various other approaches to modeling word usage and have had much success in detecting depression. Coppersmith et al. (2015) accurately identified depression with high accuracies using n-gram models in Twitter text. Althoff et al. (2016) presented a large-scale quantitative study on the discourse of counseling conversations. They developed a set of discourse features to measure how correlated linguistic aspects of conversations were with outcomes. They used a dataset of approximately 80,000 text message counseling conversations. On average, each conversation lasted about 40 messages and after the conversation ended texters received a follow-up question of "How are you feeling now?", which they use as their ground-truth label. Features in their study included: sequence-based conversation models, language model comparisons, message clustering, and psycholinguistics-inspired word frequency analyses. Their results were also consistent with Psyzczynski and Greenberg's theory of depression, in that texters with a smaller amount of self-focus were associated with more successful conversations. In addition, Schwartz et al. (2014) showed that regression models based on Facebook language could be used to predict an individual's degree of depression.

In addition to considering word usage, researchers have also explored syntactic characteristics of depressed language. As discussed in Chapter 2, Section 2.2, Zinken et al. (2010) investigated whether an analysis of a depressed patient's syntax could help predict improvement of symptoms and found that certain structures were correlated with patients' potential to complete a self-help treatment. Zinken et al.'s findings demonstrate the promise in investigating syntactic characteristics of an individual's language use. Moreover, similar work has found that differences in frequencies of part-of-speech (POS) tags were useful in detecting depression from writing (Morales and Levitan, 2016b).

Resnik et al. (2015) explored the use of supervised topic models in the analysis of detecting depression from Twitter. They used 3 million tweets from about 2,000 twitter users, of whom roughly 600 self-identified as having been diagnosed with depression. This work provided a more sophisticated model for text-based feature development for detecting depression, yielding promising results using supervised Latent Dirichlet Allocation (LDA). LDA uncovers underlying structure in a collection of documents by treating each document as if it were generated as a mixture of different topics. Qualitative examples confirmed that LDA models uncovered meaningful and potentially useful latent structure for the automatic identification of important topics for depression detection.

With the rise of social media, posts on sites such as Twitter and Facebook provide an interesting domain to investigate depression. Not only do these domains provide rich text data but also social metadata which captures important social behaviors and characteristics, like number of friends/followers, number of likes, retweets, etc. De Choudhury et al. (2014) studied Facebook data shared voluntarily by 165 new mothers. Their work aimed to detect and predict onset of postpartum depression (PPD). They considered multiple behavioral features including activity (frequency of status updates, media items, and wall posts), social capital (likes and comments on status updates or media), emotional expression and linguistic style measured through LIWC. They found that experiences of PPD were best predicted by increased isolation, which was modeled by reduced social activity and interaction on Facebook and decreased access to social capital.

Wang et al. (2013) constructed a model to detect depression from online blog posts. The features they extracted included first person singular and plural pronouns, polarity of each sentence using their polarity calculation algorithm, ratio of first person singular pronouns to first person plural pronouns, use of emoticons, user interactions with others (@username mentions), and number of posts. Using 180 users, the features given above, and three different kinds of classifiers Wang et al. (2013) reported a precision of 80% when classifying between depressed and non-depressed users.

3.4.4 Multimodal Indicators

Recent research has shown the promise in using acoustic (Cummins et al., 2015b,a, 2014; Scherer et al., 2014; Williamson et al., 2014a) and visual features (Pérez Espinosa et al., 2014; Sidorov and Minker, 2014; Williamson et al., 2014a) for depression detection. Researchers have also investigated multimodal indicators for depression detection. Scherer et al. (2013a), investigated visual signals and voice quality in a multimodal system, finding that they were able to distinguish interviewees with depression from those without depression with an accuracy of 75%.

Morales and Levitan (2016b) provided a comparative investigation of speech versus text-based features for depression detection systems, finding that a multimodal system leads to the best performing system. In addition, Morales and Levitan investigated using an ASR system to automatically transcribe speech and found that text-based features generated from ASR transcripts were useful for depression detection.

Fraser et al. (2016) built a depression detection system using a large number of textual features and acoustic features. Textual features included POS tags, parse tree constituents, psycholinguistic measures, measures of complexity, vocabulary richness, and informativeness. Acoustic features include fluency measures, MFCCs, voice quality features, and measures of periodicity and symmetry. Using these multimodal features, Fraser et al. were able to detect depression with 65.8% accuracy. Related work on suicide risk assessment found that multimodal indicators were able to discriminate between suicidal and non-suicidal patients (Venek et al., 2016).

3.5 Evaluation

Depression detection can be divided into three different prediction tasks: presence (depressed vs. not depressed), severity (normal, mild, moderate, severe, and very severe), and score level prediction. With each task comes a set of evaluation metrics. In regards to the first two groups, performance is usually reported in terms of classification accuracy (Acc.). Given that accuracy is heavily affected by skewness in datasets, often times recall or sensitivity (Sens.), specificity (Spec.), precision (Prec.), and F1-score (harmonic mean of precision and recall) are also reported. For score level prediction, performance is usually reported as a measure of differences between values predicted and the values actually observed, such as Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). In Table 3.2 we report, to our knowledge, the best performing depression detection systems from 2016.

Table 3.2: Best performing depression detection systems from 2016.

Reference	Task	Features	MAE	Acc.	Spec.	Sens.	Prec.	F1
(Fraser et al., 2016)	Binary	MFCCs, lexical, and syntax		0.66	0.61	0.71		
(Milne et al., 2016)	4 classes	N-grams		0.78				
(Kim et al., 2016)	4 classes	TF and post embedding		0.85				
(Malmasi et al., 2016)	4 classes	Lexical, syntax, and metadata		0.83				
(Brew, 2016)	4 classes	TF and metadata		0.79				
(Valstar et al., 2016b)	Binary	Visual				0.78	0.47	0.58
		Acoustic				0.89	0.27	0.41
		All				0.78	0.47	0.58
	PHQ-8	Visual	6.12					
		Acoustic	5.72					
		All	5.66					
(Williamson et al., 2016)	PHQ-8	Visual	5.33					0.53
		Acoustic	5.32					0.57
		Semantic	3.34					0.84
		All	4.18					0.81
(Yang et al., 2016)	PHQ-8	Visual and acoustic	6.70			0.67	0.50	0.57

Note. F1-score, precision, and sensitivity are reported for the depressed class. TF represents TF-IDF n-grams.

As Table 3.2 highlights, it is very difficult to make systematic comparisons across studies; data, task, label, and experimental set-up tend to vary across study. Therefore, it is hard to understand which approach is most promising. However, in regards to features, it tends to be the case that combining features from multiple modalities leads to improvements (Fraser et al., 2016; Morales and Levitan, 2016a; Scherer et al., 2013a; Valstar et al., 2016b; Williamson et al., 2016). In many cases, researchers may only have access to certain labels. However, when data sources do contain score labels reporting both error for regression as well as classification performance metrics will help facilitate comparisons across systems. Given that each feature or subset of features are meant to measure specific depression indicators or symptoms, it is also extremely important to understand how well each feature is performing. Therefore, it is best to always include correlation experiments, such as Pearson correlation tests, in order to make it transparent which features are important.

Currently, state-of-the-art performance, in terms of accuracy or error, is considered the benchmark for a well performing system. However, a discussion is needed regarding how these metrics should be used in practice. In addition, a better understanding is needed of which metrics should be considered most important, and subsequently, an agreement should be made of what the ideal performance should be. For example, how should a depression detection system report its findings to a clinician? Would the clinician prefer a BDI-II score, a binary label (depressed or not), a multi-class label (low, medium, or high), or perhaps a score with confidence intervals? Similarly, what level of performance is acceptable? For example, it may be the case, that perfect recall, to ensure no depressed cases are missed, would be the ideal performance. An alternative could be a very low error which is unlikely to misclassify a person into a different severity level. There are a wide range of possibilities. In order to determine how best to evaluate and report performance more communication between researchers is necessary, especially between those building the systems and clinicians.

3.5.1 Confounding Factors

Specific variability factors have been shown to be strong confounding factors for depression detection systems (Cummins et al., 2011, 2013, 2014, 2015a; Sturim et al., 2011). Variability factors include traits like gender, age, emotion, or personality of the speaker. Therefore, it is important to keep these factors in mind when building a detection system. For example, in many studies, systems have achieved better results using sex-dependent classifiers (Low et al., 2011; Moore et al., 2008; Scherer et al., 2014; Yang et al., 2016). Others (Morales and Levitan, 2016a) have used unsupervised clustering prior to depression detection, finding that this approach could tease out participant differences and in turn lead to performance improvements. However, these approaches to dealing with variability factors usually mean a reduction in training data, which at times can be a substantial trade-off.

Another factor to consider, is comorbidity. Comorbidity refers to the simultaneous presence of two chronic diseases or conditions. For example, Alzheimer's disease (AD) and depression frequently co-occur. Fraser et al. (2016) found that their depression detection system performed considerably lower on patients with comorbid depression and AD than on those patients with only depression. Therefore, comorbidity can lead to a more difficult task given the wide overlap of symptoms in the two conditions. Factors such as gender, age, and comorbidity, can have substantial effects on system performance. In order to better understand performance across studies and the effect of variability factors more transparency is necessary, in regards to dataset details and descriptions. In addition, researchers should begin to consider more diverse populations in their studies. Thus far, most research and data collection efforts have focused on detecting depression from young and otherwise healthy participants. In order to generalize detection systems, datasets representing other populations need to be considered.

3.6 Discussion

In this chapter, we present a review of the latest work on depression detection systems. We provide a cross-modal review of indicators for depression detection systems, covering visual, acoustic, linguistic, and social features. We also outline approaches to defining and annotating depression, existing data sources, and how to evaluate depression detection systems. This work serves as a bridge between the subfields by providing the first review of depression detection systems across subfields and modalities. Given that depression detection is inherently a multimodal problem, this work is an important contribution to the research community as it serves as a great resource for understanding multimodal features as well as what factors to consider when designing a depression detection system.

Chapter 4

A Comparative Analysis of Features for Depression Detection Systems

4.1 Motivation

In this chapter we present published work[1]. This chapter focuses on feature development for depression detection systems by investigating how to build a detection system that extracts features from multiple linguistic signals. This work aims to discover which features provide the best discrimination between depression levels. As discussed in Chapter 2, Section 2.2, the entire set of operations by which a speaker transforms ideas into acoustic output is enormously complex (Cooper and Paccia-Cooper, 1980; Lenneberg et al., 1967). When a speaker plans to produce a meaningful utterance, information is processed at a number of different levels. Considering the complexity of the speech production process, it is clear that many stages of the process can ultimately affect a speaker's acoustic output. We know that the stages of semantics and syntax affect speech (Cooper and Paccia-Cooper, 1980). We also know that a multitude of psychological symptoms can be assessed by the analysis of language

[1] Morales, M. R. and Levitan, R. (2016b). Speech vs. text: A comparative analysis of features for depression detection systems. In *IEEE Workshop on Spoken Language Technologies.*

behavior (Blanken et al., 1993). Specifically for individuals with depression, many linguistic variables have been shown to be affected, including prosody (Hönig et al., 2014; Mundt et al., 2012; Trevino et al., 2011; Yang et al., 2013), syntax (Zinken et al., 2010), and semantics (Oxman et al., 1988; Rude et al., 2004). Therefore, we assume that when we build a depression detection system we should develop features motivated from as many stages of the speech production system as possible. This chapter investigates this hypothesis. We use a subset of the AVEC 2014 corpus, which represents spontaneous speech from depressed/non-depressed individuals (Valstar et al., 2014). From this data we extract speech-based acoustic features meant to capture prosody. On the same corpus, we then perform ASR to generate transcripts of the audio. From these transcripts, we extract both text-based syntactic (structure) and semantic (content) features. To the best of our knowledge, this work presented one of the first systematic comparisons of automated text-based features with speech-based features for depression detection evaluated on the same corpus. In addition, this work was the first to explore using ASR output instead of manual transcriptions to derive text-based features, providing an evaluation of how a fully automated depression detection system may perform. We additionally provide an evaluation of speech-rate measures, derived from both speech and text at different levels (word/syllable/phoneme), determining which speech rate measure is most effective in capturing differences between depression levels.

The research questions and subsequent novel contributions of this work are the following:

1. How do prosodic features compare to text-based features for depression detection as measured by MAE/RMSE?
2. Accordingly, is incorporating ASR into a depression detection system pipeline a worthwhile pursuit?
3. How do features based on content compare to features based on structure as measured by MAE/RMSE?
4. Which speech-rate feature(s) is most effective in capturing differences between depres-

sion level?

5. Which feature(s) are most correlated with depression level?

4.2 Related Work

As discussed in Chapters 2 and 3, research has shown the promise in using speech as a diagnostic and monitoring aid for depression (Cummins et al., 2015b,a, 2014; Scherer et al., 2014; Williamson et al., 2014a). Depressed speech has consistently been associated with a wide range of prosodic, source, formant and spectral features. Consequently, the effects of depression reflected in the speech production system make speech a feasible candidate for an objective marker of depression. Moreover, studies have found that language, including syntax and semantics, is affected by depression (Morales et al., 2017a; Oxman et al., 1988; Rude et al., 2004; Zinken et al., 2010). For a thorough review of depression detection systems from speech and language, see Chapter 3, Sections 3.4.2 and 3.4.3.

4.3 Dataset

In this work, we use the AVEC 2014 corpus (Valstar et al., 2014). In total, the corpus includes 300 videos in German; the duration ranges from 6 seconds to 4 minutes. The corpus includes a total of 84 subjects. The mean age of subjects is 31.5 years, with a standard deviation of 12.3 years, and a range of 18 to 63 years. The audio data was collected using a headset microphone connected to the sound card of a laptop and sampled at various sampling rates. The corpus was re-sampled to 16 kHz. Since we are concerned with spontaneous language, we only use half of the corpus from the spontaneous *'freeform'* speech task. In total, this subset of the corpus is composed of 150 audio recordings. The task asked participants to respond to one of a number of questions such as: *What is your favorite dish or discuss a sad*

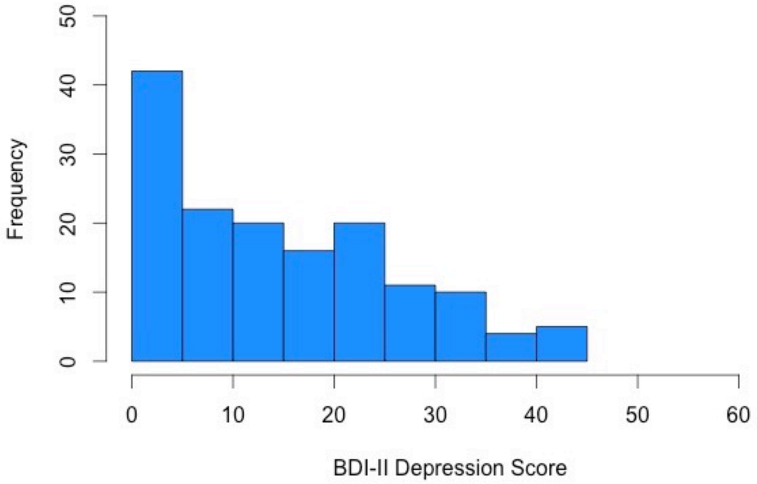


Figure 4.1: Distribution of BDI-II depression scores for the spontaneous speech subset of the AVEC 2014 corpus.

childhood memory, etc.[2]. Each recording is labeled for severity of depression. Depression severity is determined using the Beck Depression Inventory-II (BDI-II), which is the 1996 revised version of the original BDI (Beck et al., 1961). Each item of the BDI-II is a multiple-choice question scored on a discrete scale with values ranging from 0 to 3. Final BDI-II scores range from 0 to 63 (0-13 no or minimal depression, 14-19 mild depression, 20-28 moderate depression, 29-63 severe depression).

4.4 Features

4.4.1 Speech-Based Features

The speech-based features we generate include prosody features (F_0, voicing probability, loudness contours) and speech-rate features. These features were chosen to capture prosody and we base this choice on previous work which found these features useful for depression

[2] Important to note, the type of question a participant received would not only affect the topic of words chosen, but possibly also elicit different emotional responses. Therefore, responses should ideally be split by question topic prior to training/testing a detection system.

detection (Hönig et al., 2014; Mundt et al., 2012; Yang et al., 2013). F_0, voicing probability and loudness contours are extracted using the OpenSmile toolkit (Eyben et al., 2013). These features are extracted at each 10 millisecond frame. Following other work that has used OpenSmile features for depression detection and emotion recognition, we applied statistical functions to these features: arithmetic mean, root quadratic mean, standard deviation, maximum, minimum, skewness, kurtosis, quartiles, interquartile ranges, 1% percentile, 99% percentile, percentile range 1%–99%, and percentage of frames loudness contour is above: minimum+25%, 50%, and 90% of the range.

We extract a number of different speech-rate features from the audio and transcripts. Speech-rate measures derived directly from the audio include, syllable rate, average syllable duration, total speech time, total pause time, average pause time, average phone duration, and total duration. In order to extract syllable rate and average syllable duration from the audio, we use the tool AuToBI (Rosenberg, 2010) to generate pseudosyllable hypotheses. Syllable rate is defined as:

$$SyllableRate = \frac{number\ of\ syllables}{duration\ in\ seconds}$$

The AuToBI syllabifier tool is based on a procedure for automatic blind syllable segmentation for continuous speech, described in (Villing et al., 2004). Syllable rate and average syllable duration are computed using the pseudosyllables and the duration output from AuToBI . The remaining speech-rate features are generated using the BUT phoneme recognition tool (PhnRec; Schwarz, 2009). PhnRec supports four languages: Czech, English, Hungarian and Russian. We know that German is not represented and we also know that the sounds in one language may not always occur in another language. However, research in the field of language identification has found that a multilingual approach to PRLM (Phone Recognition and Language Modeling) is much better than any of the language specific PRLMs for automatically detecting languages (Ferrer et al.). This finding suggests that having coverage of information from multiple languages is more useful than having information from a single matched language. Therefore, we hypothesize that using phone hypotheses from multiple languages could be useful when generating speech-rate measures for our German data. We generate four language phone hypotheses using the trained models in PhnRec. The output of the PhnRec tool consists of phone/pause hypotheses, duration and confidence scores. Using this output, we calculate total speech time, total pause time, average pause time, average phone duration, and total duration. Our final speech-based features are comprised of the speech-rate features and the statistical functions applied to F_0, voicing probability, and loudness contours.

4.4.2 Text-Based Features

Since the AVEC 2014 corpus did not include transcripts, in order to extract text-based features, ASR was performed on the corpus. Google's German Web Speech API was used. Of the 150 audio files that comprise the corpus, 19 files received no output from the ASR. These files were then reviewed; 7 of the files were found to contain no speech audio. Although the remaining 12 files contained speech, the audio was either very low in volume or very noisy, which likely caused an issue for the ASR. The 7 non-speech audio files were not included in subsequent experiments. Important to note, the ASR output includes no information about capitalization, punctuation, or sentence boundaries. Using the transcript[3], text-based features were generated. Specifically, we borrow from previous work that has explored the relationship between an individual's writing and depression severity (Rude et al., 2004; Zinken et al., 2010). Our text-based features are comprised of two types of features, structure (or syntactic) features and content (or semantic) features.

[3] ASR transcripts will be available per request to other researchers working on this task who are licensed to use the AVEC corpora.

The words we use reflect who we are, how we feel, and what social relationships we are in. In order to access that information we generate **content features**, which represent lexical (word) choice. Our content features are based on the German version of LIWC (Pennebaker et al., 2007). As mentioned previously, LIWC is a text analysis tool that can be used to count words in psychologically meaningful categories (Tausczik and Pennebaker, 2010). The German version of LIWC is based on the English 2001 LIWC dictionary, which was translated into German. The creators of the German version of LIWC have validated this version showing that German LIWC categories display high equivalence to their English counterparts (Wolf et al., 2008). LIWC categories provide a way to capture the semantic content of the language produced. LIWC categories that are of special interest to this task include positive vs. negative emotion words, words referencing family/friends/society, pronouns which can capture inclusive language (*us, we*) vs. exclusive language (*you, they, them*), and words referencing how the person is feeling (*sad, anxious, sleep*).

In addition to LIWC features, we also built word and character level n-gram features. We take each transcript and generate n-grams from the word and character level. For the word level we consider unigrams up to trigrams. We hypothesize that unigrams would be helpful in capturing topic, whereas bigrams (2 consecutive words) and the other larger n-grams would be helpful in capturing phrases (noun/verb/adverbial phrase) as well as word order information. For the character level, we consider unigrams up to 5-grams. For word n-grams, all words are lower-cased and then stemmed because we do not wish for different conjugations of verbs to count as separate words or for plural/singular variation of nouns to count as separate words. We use the German snowball stemmer provided in Weka. However, when we build features from the character level no stemmer is used because we wish to maintain morphological endings. In order to find the balance between dimensionality size of the feature vector and coverage, we experimented with different limits for the number word/character n-grams kept; in increments of 500, we explored maxima of 500 to 3,000.

Preliminary experiments demonstrated that performance of these features for depression detection peaked around 2,000. Accordingly, all reported experiments use the threshold of 2,000 n-grams. For each n-gram, we encode presence/absence (not frequency). Therefore, for a given transcript the word/character n-gram features represent a 2,000-dimensional feature vector, which encodes the presence/absence of each given n-gram feature.

Our **structure features** include POS tag n-grams and text-based speech-rate features. In order to generate POS tags, we used the Stanford Parser toolkit which includes a German tagger (Rafferty and Manning, 2008). The tagger was trained on the Negra corpus (Skut et al., 1997) and uses the Stuttgart-Tübingen Tagset (STTS). The STTS consists of 54 German POS tags. Similar to the word and character n-gram features, using the string of POS tags from the transcript, we generate POS n-grams. We consider unigrams, bigrams, and trigrams. Our final feature set represents 2,000 mixed (unigram-trigram) POS n-grams. For each transcript, we have a 2,000-dimensional feature vector, which encodes the presence/absence of each POS n-gram. We note that POS tags represent the most simplified form of structure; future work will include transcript and audio alignment in order to capture higher level syntactic features, such as sentence boundaries or parse tree structures. In addition to POS tag n-grams, we also generate speech-rate features from the transcripts. These text-based speech-rate features include the total number of words, the total number of characters, the total number of syllables, and the average word length. We determine the number of syllables using the 'Pronouncing' package in Python, which uses the CMU dictionary (Weide, 1998) to count the number of syllables in a word. Although the CMU dictionary is meant for English, our results using these features on German speech provide a lower bound on their potential utility.

4.5 Results

Since depression is measured on a severity scale, this task represents a single regression problem. For ease of comparison to other work on this corpus, we chose to adopt the evaluation metrics of the AVEC 2014 Depression Sub-challenge: MAE and RMSE (Valstar et al., 2014). Several statistical metrics can be used to evaluate regression model performance; RMSE and MAE are widely used. We report both measures because the AVEC 2014 Depression Sub-challenge uses both and because there exists no consensus on the most appropriate metric for model errors (Chai and Draxler, 2014).

In order to put our results in context with related work, we report the results of the AVEC 2014 challenge baseline and challenge winner's performance in Table 4.1. However, differences do exist between our system and the systems that competed in the AVEC 2014 challenge, largest among them being the amount of data used in this work versus the challenge. Since our system is targeting spontaneous language, we can only use half of the AVEC 2014 corpus. In addition, we do not incorporate any visual-based features. Although we cannot make any direct comparisons, it is still useful to see how our systems compare to the state-of-the art.

Table 4.1: Depression detection baseline systems.

Features	MAE	RMSE
AVEC 2014 audio baseline	10.04	12.57
AVEC 2014 video baseline	8.86	10.86
MIT-Lincoln Challenge Winner Audio/Video System	6.31	8.12

Note. Performance reported for the AVEC 2014 challenge baseline systems (Valstar et al., 2014) as well as the challenge winner (Williamson et al., 2014b). These results were reported on the AVEC 2014's held-out test set, n=50.

To evaluate the performance of each feature set, we run leave-one-out cross-validation using SVM regression in Weka (complexity parameter C=.01) (Hall et al., 2009). SVM approaches have been shown to perform well on this task (Cummins et al., 2015a).

First, each feature set was tested in isolation. Feature sets were then aggregated into groups based on which signal they were generated from, i.e. speech or text. We also consider a combined system based on speech and text. Lastly, using the combined system, we perform feature selection to reduce the size of the feature set and to target the best performing features. We use the best first search correlation-based feature subset selection method in Weka (Hall, 1999). This method evaluates the worth of a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy between them. Using feature selection, we reduced the combined system into the best selection of features. Given the output of predictions for each model, we also calculate R^2, which indicates the proportion of the variance accounted for by each feature set. In other words, R^2 provides a measure of how well the observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model (Anderson-Sprecher, 1994).

The result of each of feature set is given in Table 4.2. The results show that our best performing feature set is the best selection model with an MAE and RMSE of 7.56 and 9.21 respectively. In regards to how a speech-based system performs when compared with a text-based system, we find that in terms of RMSE and R^2 , the text-based system performs better.

Although the speech-based system reports a lower MAE, this is likely due to the fact MAE is less sensitive to the occasional large error because it does not square the errors in the calculation. Differences between the two systems (speech-based and text-based) are not significant. The results also show that combining text and speech-based features leads to better results than using them in isolation, supporting our hypothesis that when building depression detection systems researchers should consider both signals. In addition, Table 4.3 gives a breakdown of the number of features per type for the best selection model clearly

Table 4.2: Performance by feature set.

Features	MAE	RMSE	R^2	# Features
Combined	8.91	10.72	.057	6,895
Best selection	**7.56**	**9.21**	**.238**	88
Speech-based	8.59	10.70	.064	57
Text-based	8.99	10.75	.055	6,838
Speech-rate (speech & text)	8.82	10.90	.044	26
Prosody	8.77	10.82	.038	38
Speech-rate (speech-based)	8.80	10.83	.060	19
Speech-rate (text-based)	9.10	11.09	.005	7
LIWC	9.16	11.03	.005	68
Word n-grams	9.27	11.13	.002	2,526
Char. n-grams	9.03	11.01	.060	2,098
POS n-grams	9.43	11.49	.001	2,139

Note. Results of leave-one-out cross-validation (n=138) using SMO regression (complexity parameter C = 0.01).

showing the value in text-based features. We find that a fully automated text-based system reports an MAE and RMSE of 8.99 and 10.75, which is better than many existing depression detection systems. For example, in terms of RMSE the text-based system outperforms both the audio and visual AVEC 2014 Challenge baselines. These results show the value in incorporating an ASR into a depression detection system. Future work will compare an automated text-based system to a text-based system generated from manual transcripts to investigate how a fully automated system compares.

Table 4.3: Number of features per type for the best selection feature set.

Description	Count
Prosody	1
Speech-rate	2
LIWC	2
POS n-grams	39
Word n-grams	40
Character n-grams	4

In regards to the differences between types of text-based features, we note that features which target content perform better, although the difference in performance is slight, which does not rule out the potential of syntactic features. Moreover, Table 4.3 shows that a substantial number of POS n-grams were selected as very predictive features and were subsequently used in the best selection model, which shows the promise in structural features. In terms of what level is more important when building content n-grams, character n-grams perform better than word n-grams, suggesting that individual characters/phonemes or morphological endings are more important than lexical choice.

Table 4.4: Results from statistical analyses of features.

Feature	Spearman Correlation ρ
F_0 mean	-.233**
F_0 range	-.237**
F_0 99% percentile	-.237**
F_0 standard deviation	-.246**
F_0 root quadratic mean	-.248**
Voicing 1% percentile	.190*
Loudness 1% percentile	-.290***
RU total speech	-.182*
HU total speech	-.201*
CZ total speech	-.242**
HU avg. phone duration	-.269**
CZ avg. phone duration	-.322***
Duration	-.193*
Job	.18**
Pronoun (Other)	.303***
Posfeel	.169*
Sad	.188*
Sleep	.237**
You	.223**

Note. Features included in these analyses included only prosodic features, speech-rate features, and LIWC features (*p<.05; **p<.01; ***p<.001).

In addition to cross-validation experiments, we also ran statistical analyses for all features to discover which feature(s) were highly correlated with depression level. The results for the statistical analyses of features are reported in Table 4.4. Spearman's ρ was computed between each individual feature and depression level. Table 4.4 reports only the features that correlated significantly with depression. We find a number of prosodic features to be negatively correlated with depression level. In particular, a number of F_0 features are correlated with depression level, supporting previous research that F_0 is affected by depression level (Blanken et al., 1993; Cummins et al., 2015a). Specifically we find F_0 mean, F_0 standard deviation, and F_0 range to all be significantly correlated with depression level. We also find many speech-rate features to be correlated with depression level. Interestingly, speech-rate measures which measure the total speech produced or the average phone duration are most correlated with the BDI-II score; surprisingly we find features that capture total pause time or average pause duration were not significant. These results also show the promise in text-based features, providing evidence that text-based features have potential, confirming previous findings (Valderas et al., 2009). In particular, we see a number of LIWC categories are positively correlated with depression level. Table 4.5 provides (translated) examples of words in each of the LIWC categories.

As previous work has found (Rude et al., 2004; Stirman and Pennebaker, 2001), pronoun use and negatively valenced words such as the words in the sad category are correlated with depression level. We also find that depression level is correlated with word use related to labor/occupation and sleep; these findings are especially interesting, since words from these two LIWC categories can closely be associated with language related to many depression symptoms, including fatigue, loss of energy, insomnia and hypersomnia (APA, 2013). Since the corpus of spontaneous speech used for these experiments included participants' responses to a number of different questions, it would be interesting to see whether or not there is

Table 4.5: Examples of words in the LIWC categories that are strongly correlated with depression severity score.

LIWC Category	Examples
Job	job, labor, manag*, market
Pronoun (Other)	he, him, her, hers
Posfeel	admir*, grin, happy, joy
Sad	crying, grief, sad, sorrow
Sleep	asleep, awake, bed, daze*
You	you, your, you're, you'll

Note. Words were translated from German into English. * represents the morphological stem of the word.

a performance difference for different questions. Future work will include an analysis by question type.

Lastly, we were interested in how accurate our ASR transcripts were. We had a native speaker of German transcribe[4] all the audio files. We then calculated the word error rate (Klakow and Peters, 2002): 57.05%. Given the current state-of-art performance measures for ASR systems, this represents a very high error rate. When we looked further into what type of errors (additions, substitutions, deletions) were found, we found that the ASR transcripts were mostly missing a lot of content. However, of the words the ASR transcribed 70% were accurate. Although ASR performance is poor, our models still perform well, which leads us to believe ASR can still be a valuable contributor to the depression detection pipeline. Future work will investigate the differences between how these features perform with ASR output versus manual transcriptions to determine how well an ASR needs to perform in

[4] Manual transcripts will be available per request to other researchers working on this task who are licensed to use the AVEC corpora.

order to be included in a system.

4.6 Discussion

This work provides a comparative analysis of features for depression detection. To our knowledge, this work presented the first ever automated speech to text-based system for depression detection. Using the same corpus, this work also presented a systematic comparison of a text-based system for depression detection with a speech-based system. We find that a small set of 88 hypothesis-driven linguistically-motivated features derived from both the speech and text signal perform very well, reporting an MAE and RMSE of 7.56 and 9.21. As a consequence, we stress the value of incorporating an ASR into the depression detection system pipeline. There exist many stages to the speech production process and researchers should draw from as many stages as possible, including the phonetic, semantic, and syntactic level. We also find strong correlations between word use and depression level. We provide support for previous findings that found pronoun use and negatively valenced words to be correlated with depression. We also provide new evidence that a relationship exists between language related to work/sleep and depression level.

Chapter 5

Multimodal Approaches to Detection

5.1 Motivation

In this chapter we present published work[1]. When focusing on classification tasks, researchers across various fields—speech processing, NLP, and HCI—often work on the same task from different perspectives, usually with different data sources and feature representations. However, to truly get a comprehensive picture of a conversation, it is necessary to consider all modalities. In many situations, a multimodal system can provide the most robust source of information for a classification task and research has found that on average multimodal systems offer an 8% improvement over unimodal systems (D'Mello and Kory, 2012). However, building a multimodal system is extremely time intensive because it requires feature engineering across multiple modalities. In some cases, it is also not feasible given the dataset type. This work presents a novel multimodal feature extraction tool: **OpenMM**. The goal of OpenMM is to provide researchers with a simple tool to extract multimodal features. OpenMM is built upon various existing open-source tools as well as our own code for linguistic analysis. The tool only requires a video as input and performs all the processing and

[1] Morales, M. R., Scherer, S., and Levitan, R. (2017b). OpenMM: An Open-Source Multimodal Feature Extraction Tool. In *Proceedings of Interspeech 2017*, pages 3354-3358, Stockholm, Sweden. ISCA.

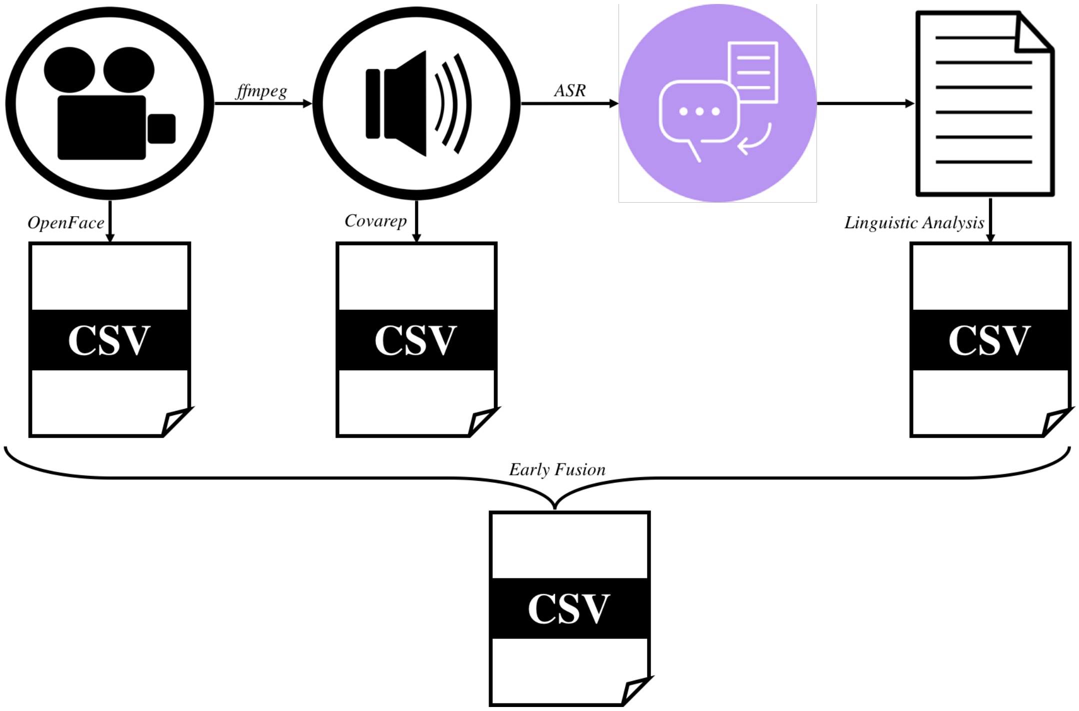


Figure 5.1: OpenMM Pipeline

necessary conversions to generate audio files and transcriptions, as shown in Figure 5.1.

Given a video input, OpenMM will extract visual features using the open-source tool OpenFace (Amos et al., 2016). Then OpenMM converts the video to audio using the tool *ffmpeg* (Tomar, 2006), outputting an audio wav file. Using the wav file, OpenMM extracts acoustic features using the open-source repository Covarep (Degottex et al., 2014). Using the wav file, OpenMM then makes a call, depending on the language, to either IBM Watson's or Google's speech-to-text service, outputting a transcript. Using the transcript OpenMM then extracts linguistic features, which include bag-of-words features and syntactic features. The syntactic features are generated using a dependency parse tree representation, which is generated using Google's state-of-the-art parser. In the end, OpenMM outputs the following: wav file, transcript file, comma-separated values file (CSV) of visual features, CSV of acoustic features, CSV of linguistic features, and CSV of multimodal features. OpenMM currently supports English, German, and Spanish and is available for download[2].

We evaluate the OpenMM multimodal feature set on three different classification tasks. In addition to depression detection, we choose to incorporate related affect detection tasks to test the robustness of OpenMM's feature set. The three classification tasks we consider are depression detection (depressed vs. not depressed), deception detection (deceptive vs. truthful), and sentiment detection (negative vs. positive), which respectively involve datasets in English, German, and Spanish. In many experiments, we find OpenMM features match or outperform state-of-the-art systems (Pérez-Rosas et al., 2013, 2015). Using OpenMM we are able to classify depression with 76.79% accuracy, deception with 76.86% accuracy, and sentiment with 62.50% accuracy. We hope this tool will provide researchers with a simple and inexpensive way of extracting multimodal features.

5.2 Related Work

In this section we provide brief overviews of related work on the three relevant classification tasks.

5.2.1 Deception Detection

Pérez-Rosas et al. (2015) presented the first multimodal system to detect deception in real-life trial data using text and gesture modalities. They built classifiers relying on individual and combined sets of nonverbal and verbal features, reporting accuracies in the range of 60-75%. Their dataset was manually transcribed and their verbal features included unigrams and bigrams derived from the bag-of-words representation of their video transcripts. Their nonverbal features included facial displays and hand gestures, which were also manually annotated. Pérez-Rosas et al.'s findings showed the promise in multimodal features while

[2] `https://github.com/michellemorales/OpenMM`

also highlighting the time intensive nature of multimodal design, which often includes a good deal of manual annotation.

5.2.2 Sentiment Detection

In other work, Pérez-Rosas et al. (2013) presented a method for multimodal sentiment classification, which could identify the sentiment of video reviews. In order to identify sentiment, they explored visual, acoustic, and text features. Acoustic features were extracted using the open-source software OpenEAR (Eyben et al., 2009), including prosody, energy, voice probabilities, spectrum, and cepstral features. Facial features were extracted using the Computer Expression Recognition Toolbox (Littlewort et al., 2011), including smile and head pose estimates, facial AUs, and eight basic emotions. Lastly, video clips were manually transcribed. In addition, linguistic features were extracted including a bag-of-words representation. Their work showed that multimodal sentiment analysis can be effectively performed. They also report that the joint use of multimodal features (visual, acoustic, and linguistic) can lead to error rate reductions of up to 10.5% as compared to the best performing single modality.

5.2.3 Depression Detection

As discussed in Chapter 3, researchers have also investigated the use of multimodal features for depression detection. Recent research has shown the promise in using acoustic features (Cummins et al., 2014, 2015a,b; Scherer et al., 2014; Williamson et al., 2014a) and visual features (Pérez Espinosa et al., 2014; Sidorov and Minker, 2014; Williamson et al., 2014a) for depression detection. Some researchers have built multimodal systems, specifically Scherer et al. (2013a), who investigated visual signals and voice quality, finding that they were able to distinguish interviewees with depression from those without depression with an accuracy of 75%. In addition to audiovisual features, text-based features, including syntactic and semantic features have been investigated for depression detection. As mentioned previously, Rude et al. (2004) examined linguistic patterns of student narratives, finding that depressed students used significantly more first person singular words and negatively valenced words than did never-depressed students. In addition, Zinken et al. (2010) investigated whether an analysis of a depressed patient's syntax could help predict improvement of symptoms and found that certain syntactic structures were correlated with patients' potential to complete self-help treatment.

5.3 Datasets

In these experiments, we use three publicly-available datasets: the Real-life Trial (RLT) dataset[3], the Multimodal Opinion Utterances dataset (MOUD), and the AVEC 2014 dataset[4].

5.3.1 Real-life Trial Dataset

The RLT dataset (Pérez-Rosas et al., 2015) includes videos of real deception during court trials in English. Videos were collected from public multimedia resources where trial hearing recordings were available and where truthful or deceptive behavior could be fairly observed and verified. Videos selected met the following guidelines: defendant or witness in the video should be clearly identified, his/her face should be visible during most of the clip duration, visual quality should be clear enough to identify facial expressions, and clear audio quality. Trial outcomes, such as guilty verdict, non-guilty verdict and exoneration, are used to help correctly label video clips with deceptive or truthful.

[3] The MOUD and the RLT datasets can be found here: `http://web.eecs.umich.edu/~mihalcea/downloads.html`

[4] `https://avec2013-db.sspnet.eu/`

5.3.2 Multimodal Opinion Utterances Dataset

The dataset we use for sentiment classification is the MOUD dataset (Pérez-Rosas et al., 2013). The MOUD dataset includes videos of product opinions expressed in Spanish. The videos were collected from the social media web site YouTube, using several search keywords that were likely to lead to product reviews or recommendations. Videos selected met the following guidelines: the speaker was directly in front of the camera, his/her face was clearly visible, with minimum amount of face occlusion, and no background noise. In total the dataset is comprised of 80 videos randomly selected from the videos retrieved from YouTube that met the guidelines. All video clips were manually processed to transcribe the verbal statements and to extract the start and end time of each utterance. Each utterance was then labeled for sentiment by two annotators.

5.3.3 Audio-Visual Emotion Recognition Challenge Dataset

For the depression classification task, we use the AVEC 2014 corpus (Valstar et al., 2014). In total, the corpus includes 300 videos in German. Since we are concerned with spontaneous language, we only only use half of the corpus from the spontaneous *'freeform'* speech task. In total, this subset of the corpus is composed of 150 videos. The videos include recordings of participants responding to one of a number of questions. Each recording is labeled for severity of depression using the BDI-II (Beck et al., 1961). BDI-II scores range from 0 to 63. We group the data into 2 binary classes not depressed (0-13) and depressed (≥14).

5.4 OpenMM Feature Extraction

OpenMM aims to extract features from as many channels or modalities as possible, including nonverbal behavior, voice and speech characteristics, as well as linguistic characteristics.

5.4.1 Automatic Speech Recognition

Given advancements in ASR, language can now be a common component in classification systems. For this reason, it is important to investigate how successful a feature set can be when it is fully automated. Manual transcription ensures the most accurate transcription possible, however it is expensive in time and resources. Therefore, it is important to investigate how ASR transcript derived features compare to manual transcription derived features. OpenMM includes ASR to automate the transcription process. For English and Spanish, we use Watson's Speech-to-Text API [5]. For German, we use Google's API [6].

5.4.2 Verbal Features

Bag-of-words

Each ASR transcript represents a string of words, with no punctuation or capitalization included. We take each transcript and generate a bag-of-words representation of each sentence to derive unigram counts, which are then used as linguistic features. We first build a vocabulary consisting of all the words occurring in the transcriptions. We then remove words that have a frequency below 10. This threshold is based off previous work which found this threshold useful for deception and sentiment detection (Pérez-Rosas et al., 2013, 2015). The remaining words represent the unigram features. So for each sentence, we generate a feature vector that represents the frequency of the unigrams inside that utterance.

Syntax Features

In order to generate syntactic features, we first tag and parse all sentences, using Google's state-of-the-art pre-trained English parser: Parsey McParseface (Andor et al., 2016). We also use Google's Spanish and German universal parsers. For each sentence S, the parser

[5] www.ibm.com/watson/developercloud/doc/speech-to-text/

[6] https://cloud.google.com/speech/

Table 5.1: Dependency distance measure.

Dependency Relation	Distance
NSUBJ(saw-2, I-1)	1
DOBJ(man-4, saw-2)	2
DET(man-4, the-3)	1
PREP(with-5, man-4)	1
POBJ(glasses-6, with-5)	1
Sum	6

outputs universal POS tags. Grammatical roles are also labeled, which show how words in the sentence relate to one another. For example, the sentence "*I saw the man with glasses*" when parsed would output the dependency relationships listed in Table 5.1.

Using the parser's output, syntactic features are generated, including: depth of tree, number of root dependents, number of unique universal POS tags, frequency of each POS tag, average word length, and a computed dependency distance measure. The depth of the tree represents the number of levels in the tree, which gives a measure of how complex of a construction the sentence is. The number of root dependents represents the total number of children the root has, providing another way to represent sentence complexity. The number of unique POS tags captures how many unique POS tags were used, which measures syntactic variety. The average word length represents a simple way of capturing how advanced the vocabulary is. Lastly, the dependency distance measure is based on related work (Pakhomov et al., 2011). Given each parse tree, each dependency relation receives a distance score calculated as the absolute difference between the serial positions of the words that participate in the relation, i.e. difference between indices in the sentence.

The dependency distance measure is then the sum of all the dependency distances in the sentence, as shown in Table 5.1.

5.4.3 Nonverbal Features

Facial Features

OpenFace (Baltrušaitis et al., 2016) is used to extract 408 visual features. OpenFace is an open-source facial behavior analysis toolkit, which has achieved state-of-the-art results in facial landmark detection, head pose estimation, facial AU recognition, and eye gaze estimation. OpenFace includes features that capture basic information about the video, such as frame number, timestamp, and confidence values. Features also include information about an individual's gaze as well as the location of their head and face, which are represented in the gaze, pose, and landmark features. In addition, OpenFace includes features from FACS (Ekman et al., 1978). As mentioned previously, FACS is a system used to taxonomize human facial movements by their appearance on the face. It is a commonly used tool and has become standard to systematically categorize physical expressions, which has proven very useful for psychologists. FACS is composed of facial AUs, which represent the fundamental actions of individual muscles or groups of muscles.

Acoustic Features

In order to extract features from the voice, we use Covarep: a Cooperative Voice Analysis Repository for Speech Technologies (Degottex et al., 2014). Covarep is an open-source toolkit of advanced speech processing algorithms. Using Covarep we extract 71 audio features, including prosodic, source, and spectral features. Prosodic features of fundamental frequency and voicing boundaries are extracted using a simple and robust pitch tracking algorithm (Degottex et al., 2014). Covarep features also include features derived from the glottal source signal estimated by glottal inverse filtering (Degottex et al., 2014). In addition, features include two wavelet based features and the posterior probability of the creaky voice detection algorithm included in Covarep. Lastly, spectral features include spectral envelope estimation and the mean and deviation of the harmonic model plus phase distortion.

5.4.4 Fusion

For each unimodal feature set, OpenMM outputs a CSV of features. For the visual and acoustic features, the features are computed at the frame-level. For the text-based features, features are computed at the sentence-level. In order to fuse the modalities, we need one feature vector per modality. Therefore, we apply statistical functionals to each unimodal feature set. We apply the following statistical functionals: maximum, minimum, mean, median, standard deviation, variance, kurtosis, skewness, 25% percentile, 50% percentile, and 75% percentile. Using the feature vector derived through the statistical functionals, we then fuse the modalities by concatenating each of the video-level feature vectors. In addition to the multimodal (verbal + nonverbal) feature set, we also fuse the verbal (bag-of-words + syntax) and nonverbal (acoustic + visual) modalities.

5.5 Results

We conduct three series of experiments. For each series, we build and evaluate classification models using OpenMM's feature sets.

5.5.1 Deception Detection

In order to compare directly to Pérez-Rosas et al.'s (2015) previous work on deception detection, we use the same experimental configuration. Therefore, we evaluate using two classification algorithms, Decision Trees (DT) and Random Forest (RF), using the Weka

Table 5.2: Deception accuracy reported for leave-one-out cross-validation using DT and RF algorithms.

Feature Set	DT	RF
P2015 - Bag-of-words	60.33	56.19
P2015 - Facial	70.24	76.03
OpenMM - Bag-of-words	66.94	59.50
OpenMM - Syntax	57.02	62.81
OpenMM - Acoustic	75.21	76.86
OpenMM - Visual	71.07	73.55
P2015 - Verbal	60.33	50.41
P2015 - Nonverbal	68.59	73.55
OpenMM - Verbal	61.16	59.50
OpenMM - Nonverbal	74.38	75.21
P2015 - All Features	75.20	50.41
OpenMM - All	73.55	76.03

toolkit with default parameters (Hall et al., 2009). We run several comparative experiments using leave-one-out cross-validation. In Table 5.2, we report our results in conjunction with Pérez-Rosas et al.'s results, which we refer to as **P2015**. Given the distribution between deceptive and truthful clips, the random baseline on this dataset is 50.4%. We find that the deception prediction accuracy for OpenMM's multimodal feature set is 76.03% which matches P2015's best performing system. These results are extremely promising as they confirm that OpenMM's fully automated system can match the performance of a manually handcrafted feature set. In addition, across modalities, OpenMM's performance matches or outperforms P2015's models. This is especially interesting in regards to verbal features; OpenMM's bag-of-words, syntax, and verbal feature sets outperform P2015's verbal feature sets. These findings confirm that ASR transcript derived features can compete with manual transcription derived features. Lastly, the OpenMM acoustic feature set achieves the best results, classifying deception with 76.86% accuracy.

5.5.2 Sentiment Detection

Similar to the the deception detection experiments, we also compare OpenMM's sentiment detection results directly with previous work. Results for OpenMM's models can be compared directly to Pérez-Rosas et al. (2013) systems' results, which we refer to as **P2013** in Table 5.3. As before, we evaluate each unimodal feature set as well as the multimodal feature sets. Following P2013, we use an SVM classifier in ten-fold cross-validation experiments. Given the distribution between positive and negative clips, the random baseline on this dataset is 55.93%. For sentiment detection, we find that OpenMM's unimodal acoustic and visual features outperform P2013's feature sets. However, we also find that OpenMM's verbal feature sets are unable to match the performance of P2013. We think this can be attributed to the ASR model. Specifically for Watson's speech-to-text service, IBM announced that their English conversational speech recognition system achieves an 8% word error rate. However, they also mention having little data for building the Spanish model, leading to far higher error rates (Saon, 2016). We believe this difference in ASR model performance led to poorer performing verbal feature sets. Although, we find ASR to be an extremely valuable tool for feature engineering for the English task, these results for Spanish highlight the limitations of ASR.

Table 5.3: Sentiment classification accuracy reported for ten-fold cross-validation using SMO.

Feature Set	SMO
P2013 - Verbal	73.33
P2013 - Acoustic	53.33
P2013 - Visual	50.66
OpenMM - Bag-of-words	48.96
OpenMM - Syntax	60.42
OpenMM - Acoustic	61.46
OpenMM - Visual	62.50
P2013 - Nonverbal	61.33
OpenMM - Verbal	52.08
OpenMM - Nonverbal	59.38
P2013 - All	74.66
OpenMM - All	57.29

5.5.3 Depression Detection

Lastly, we evaluate OpenMM's feature sets on the depression detection task. Results are given in Table 5.4. Given the distribution between depressed and not depressed clips, the random baseline on this dataset is 55.36%. Since we only use half of the AVEC corpus and conduct a classification experiment, instead of the more common regression, it is difficult to provide a direct system comparison for depression detection. However, given the difficult nature of the task, we believe OpenMM's results show promise. The visual, acoustic, and multimodal features perform better than the baseline. As shown in Table 5.4, the OpenMM

Table 5.4: Depression classification accuracy reported for leave-one-out cross-validation using SMO.

Feature Set	SMO
OpenMM - Bag-of-words	44.64
OpenMM - Syntax	44.64
OpenMM - Acoustic	76.79
OpenMM - Visual	62.50
OpenMM - Verbal	46.43
OpenMM - Nonverbal	62.50
OpenMM - All	62.50

nonverbal, acoustic, and visual feature sets achieve the best results. The acoustic feature set represents the highest performing system, reporting an accuracy of 76.79%. These results confirm previous findings that acoustic and visual features are extremely useful for depression detection (Scherer et al., 2013b). Similar to what we found for sentiment detection, the verbal feature sets represent the lowest performing systems, which is again likely an artifact of the German ASR model.

5.6 Discussion

In this chapter, we present OpenMM, the first open-source multimodal feature extraction tool. We evaluate OpenMM on three datasets spanning three different languages. We find that OpenMM's unimodal and multimodal feature sets perform well across different classification tasks. Our best performing models are able to classify deception with 76.86% accuracy, sentiment with 62.50% accuracy, and depression with 76.79% accuracy. Our findings show that multimodal features derived from a fully automated system can match the performance of a manually handcrafted feature set. In addition, we find that features derived from ASR transcriptions can compete with features derived from manual transcriptions. We hope OpenMM will provide researchers with a simple and inexpensive way of extracting multimodal features, which encompass various communicative modalities: face and gesture, voice and speech, and language. Lastly, we hope OpenMM can lead to richer and more robust feature representations for machine learning tasks, including depression detection.

Chapter 6

Mitigating Confounding Factors in Depression Detection Using an Unsupervised Clustering Approach

6.1 Motivation

In this chapter we present published work[1]. Speech processing researchers have investigated, at depth, speech features for depression detection (Cummins et al., 2015b). Although previous research has made great progress in understanding what acoustic features and machine learning models are most suitable for automatically predicting severity level of depression, there is a lack of exploration into dealing with sources of variability, which can significantly confound results. In general, when eliciting speech as a marker for depression the following confounding factors complicate the task: biological traits such as gender, cultural traits such as dialect, and emotional signals such as fear and anger. These variability factors place a ceiling on the accuracy of a speech based system for depression detection. Given this poten-

[1] Morales, M. R. and Levitan, R. (2016a). Mitigating confounding factors in depression detection using an unsupervised clustering approach. In *Proceedings of the 2016 Computing and Mental Health Workshop.*

tial limitation, it is important to research ways to mitigate these factors. This work presents an approach to deal with confounding factors by utilizing a two-layer architecture. To tease apart the traits/states of the speakers involved, we first perform unsupervised clustering using a K-means algorithm. We then perform depression detection on each of the clusters separately and find that clustering prior to classification can help boost performance.

6.2 Related Work

Some work has investigated mitigating confounding factors, such as speaker characteristics, phonetic content, and recording setup variability. Cummins et al. (2011) based their work on findings from emotion recognition research, hypothesizing that accurate selection of speech segments would provide maximal depressed/neutral speech discrimination (Cummins et al., 2011). They expected to find that voiced segments provided the most effective discrimination. In addition, they explored normalization techniques, such as mean and mean-variance normalization as well as feature warping, which attempts to reduce variation in data due to differences in speaker variability. They found that discriminating between voiced/voiceless speech segments was not critical to the task. Mean and mean-variance normalization techniques were not reported due to their very poor performance and feature warping as a per-speaker feature space normalization technique offered little to no improvement.

In later work, Cummins et al. (2013) provided an analysis of the AVEC 2013 speech corpus (Valstar et al., 2013). They analyzed the phonetic variability of the data by generating multiple sub-utterances per file. They then show that each sub-utterance differs vastly in phonetic content, by demonstrating that there exist a wide range of prediction scores for each file. Their analysis provided insight into the phonetic variability that exists across a depressed speech utterance.

Cummins et al. (2014) investigated acoustic volume proposing a novel GMM-based measure that is able to capture the decreasing spectral variability that is usually associated with depressed speech. Using this approach they were able to show that with increasing levels of depression the MFCC feature space narrows to become more tightly concentrated.

Some researchers have borrowed techniques from speech/speaker recognition research, which have helped mitigate confounding forms of variability. Sturim et al. (2011) were able to reduce the effects of speaker and intersession variability by using a Weiner Filtering Factor Analysis method to enhance a MFCC-based GMM system. In addition, they found that using a 2-class gender independent set up resulted in a reduction in Equal Error Rate of ~21% and ~29% for the male and female systems when compared to one single model for both genders. Sturim et al.'s findings demonstrate the influence gender differences have on a system.

Other interesting approaches have performed analyses of the relationships that exist between different symptoms of depression and different prosodic and acoustic features. Some have even found significantly stronger correlations between their measures on individual items on the HAMD, such as low mood when compared to the total HAMD score (Horwitz et al., 2013; Quatieri and Malyska, 2012; Trevino et al., 2011).

The related research discussed above serves to demonstrate the variability inherent in depressed speech as well as highlight the importance in dealing with this variability. Overall, previous work has aimed to mitigate confounding factors, by exploring different normalization techniques, statistical models, and architectures. This work builds upon previous work by introducing a multi-layer architecture, which involves unsupervised clustering. This technique is also borrowed from work in speaker identification. Clustering has proven to be a successful technique in segmenting speakers without any prior knowledge of the identities or the number of speakers (Hu et al., 2013; Kinnunen et al., 2011). Here, clustering provides a way to tease apart different sources of variability prior to depression classification.

6.3 Dataset

The data used in this work is the AVEC 2014 corpus (Valstar et al., 2014), which is a subset of the AVEC 2013 corpus (Valstar et al., 2013). For more details about the dataset see Chapter 4, Section 4.3. The AVEC 2014 corpus is already partitioned for training and development data sets. For the training and development corpora respectively the average BDI-II is 15.0 (±12.3) and 15.6 (±12.0). Each of the partitions contains 92 audio files.

6.4 Features

The feature set we use is borrowed from the AVEC 2014 baseline (Valstar et al., 2014). The set consists of 2,268 features extracted using the OpenSmile toolkit (Eyben et al., 2013). The features are composed of 32 energy and spectral related low-level descriptors (LLD) x 42 functionals, 6 voicing related LLD × 32 functionals, 32 delta coefficients of the energy/spectral LLD × 19 functionals, 6 delta coefficients of the voicing related LLD × 19 functionals, and 10 voiced/unvoiced durational features. In addition, the LLD set covers a standard range of commonly used features in audio signal analysis and emotion recognition. Features were extracted over overlapping short fixed length segments of 20 seconds which are shifted forward at a rate of one second (Valstar et al., 2014).

6.5 Clustering

Using the above mentioned feature set, unsupervised clustering was performed using the K-means clustering algorithm in Weka (Hall et al., 2009). Clustering can be defined as the unsupervised classification of patterns into groups. The resulting groups or clusters should ideally exhibit the following characteristics: (1) homogeneity within the clusters, and (2) heterogeneity between clusters. Several algorithms require certain parameters for clustering, such as the number of clusters. For K-means clustering, the number of clusters k must be specified. Since the dataset is relatively small only small values of k are explored (k=2 up to k=5). We hypothesized that low values of k would capture the most basic forms of variation, such as a gender, where higher values of k would capture more complex forms of variation. Subsequently, the clusters established were used to train different models based on each cluster. During test time, each feature vector is compared to all existing cluster centroids by computing the euclidean distance between the 2 vectors. The cluster centroid represents the average across all the points in the cluster. The closest cluster to the new feature vector in question is then chosen as the model that will be used during classification. So for example for a given cluster centroid p and a feature vector q we calculate the distance between the two using the formula below.

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}$$

We calculate the distance for every cluster centroid and the closest centroid is marked as the cluster that will be used during testing for that specific feature vector.

6.6 Evaluation

Since depression is measured on a severity scale this task represents a single regression problem. The learning algorithm we employ is SMO regression (default parameters) in Weka (Hall et al., 2009). We choose to adopt the evaluation metric of the AVEC 2014 Depression Sub-challenge: MAE (Valstar et al., 2014). Lastly, we use as our baseline, an MAE of 10.26. This MAE score is achieved by evaluating our feature set on the test data without clustering (92 instances in train and test). MAE can be defined, at a basic level, as the absolute average difference between the actual labels and the predictions made. In most work involving the AVEC 2014 corpus, RMSE is also reported; here we choose to only report MAE. We choose to adopt this baseline and not the challenge or challenge participants' baselines because our feature set is only extracted from the audio signal. Since the AVEC 2014 corpus also includes video, many systems chose to incorporate features from that signal. For this reason, a direct comparison cannot be easily made between our system and those existing systems.

6.7 Results

The results of our experiments are given in Table 6.1.

Table 6.1: MAE Results using different values of k.

# of Clusters	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Average MAE
$k = 2$	9.28	9.94	—	—	—	9.61
	t-45 T-48	t-47 T-44				
$k = 3$	17.01	10.57	7.49	—	—	11.69
	t-36 T-29	t-35 T-34	t-21 T-29			
$k = 4$	26.14	8.73	8.82	10.31	—	13.50
	t-21 T-21	t-21 T-25	t-17 T-18	t-33 T-28		
$k = 5$	23.85	10.87	8.82	10.17	5.85	11.86
	t-21 T-26	t-3 T-15	t-17 T-18	t-41 T-18	t-10 T-15	

Note. The number of clusters k ranges from $k = 2$ to $k = 5$. t represents the number of training instances and T represents the number of test instances.

The hypothesis we tested was whether or not clustering could provide a way to tease apart different sources of variability prior to depression classification. The scores achieved with clustering can be compared to our baseline: an MAE of 10.26. It can be noted that clustering in many cases does help boost performance. When the number of clusters is small we see a uniform improvement of lower MAE across clusters; cluster 1 and cluster 2 for k=2 achieve a MAE of 9.28 and 9.94 respectively. For each of the values of k we see improvements in some clusters but not in all. In some cases, we see substantially worst performance. Due to data size, it is possible that when the values of k increase performance worsens due to the small number of training instances. In other words, data size presents a limitation to this approach.

In order to make claims about what traits or labels the clusters may possibly be representing, we use two metrics to evaluate the clusters' performance. Given the knowledge of the ground truth class assignments of the data, it is possible to define some intuitive metrics. Specifically some (Rosenberg and Hirschberg, 2007) have defined the following two desirable objectives for any cluster assignment:

1. homogeneity: each cluster contains only members of a single class
2. completeness: all members of a given class are assigned to the same cluster

Scikit-learn's implementations of the above metrics are used. The labels considered are task (read speech vs. spontaneous speech), gender (male vs. female), and depression level (low/none vs. high). Levels of depression are determined by using the clinical suggestion attached to BDI-II, which suggests that an individual should seek out professional help when receiving a score of 17 and above. Therefore, any participants rated 17 and above are considered to be in the high group and the rest in the low/none group.

Scores are bounded between 0 and 1 (1 being the best). For this evaluation, only the k=2 clustered is considered. The scores for each of the metrics are given in Table 6.2. As the results show, neither of the labels received high marks. Relative to the results given, the clusters seem to be capturing task above any other trait, suggesting that this factor should be separated out prior to classification. Consequently, when task is addressed and 2 models

Table 6.2: Metrics for cluster assignment.

Label	Homogeneity	Completeness
Task	.0584	.0581
Gender	.0002	.0002
Depression level	.0068	.0065

Note. Depression level represents a discrete label of low versus high.

are trained (one for each task), the MAE results improve to 7.6 and 9.71 respectively for the spontaneous and read speech task. These results are consistent with previous findings that suggest task differences should be considered (Valstar et al., 2014). These metrics, to some extent, support the claim that clustering is capable of capturing variation and differences, across task, and potentially across speaker traits. Important to note, we are not very concerned with whether the clusters capture specific labels; they may be capturing any combination of factors that may affect depression detection—gender, age, accent, class, emotion, etc. We are more concerned with whether clustering helps the task of depression detection.

6.8 Discussion

There are many challenges to depression classification. This chapter focused on addressing one such challenge: variability factors. This work presented an approach based on unsupervised clustering that resulted in slight gains to performance as measured by MAE. In addition, we found that clustering prior to classification helps mitigate certain factors, such as speech task. However, as the value of k increased it seemed that data size presented as an issue. Future work should consider upsampling and data augmentation techniques to help overcome this limitation. In addition, this work only used a standard feature set with no feature analysis or development. Since cluster assignment is based solely on the features employed, feature development should be explored to improve upon current work. Lastly, only one simple clustering algorithm was explored: K-means. More sophisticated algorithms may help leverage performance gains without the drawbacks of the simple approach taken here.

Chapter 7

A Novel Fusion Approach

7.1 Motivation

Initial studies on depression detection from multimodal features have shown performance gains can be achieved by combining information from various modalities (Morales and Levitan, 2016b; Scherer et al., 2014). However, few studies have investigated, in depth, fusion approaches for depression detection (Alghowinem et al., 2015). In this chapter, we present a novel linguistically motivated approach to fusion: *syntax-informed fusion.* We compare this novel approach to early fusion and find it is able to outperform it. We also demonstrate that this approach overcomes some of the limitations of early fusion. Moreover, we test our approach's robustness by applying the same framework to generate a *visual-informed fusion* model. We find video-informed fusion also outperforms early fusion. In addition to presenting novel fusion techniques, we also evaluate existing approaches to fusion including early, late, and hybrid fusion. To the best of our knowledge, this work presents the first in-depth investigation of fusion techniques for depression detection. Lastly, we present interesting results to further support the relationship between depression and syntax.

7.2 Related Work

This work presents a multimodal detection system with a specific focus on the relationship between depression and syntax. This relationship motivates a novel approach to fusion. In contrast to a simple early fusion approach to combining modalities, a syntax-informed early fusion approach leverages the relationship between syntax and depression to help improve system performance. In this section, we first provide background on the relationship between depression and language, highlighting both the voice and syntax. In addition, we also evaluate an *informed fusion* approach which is motivated from the relationship between depression and facial activity as well as the relationship between facial behavior and speech production. Therefore, we also present related work on the relationship between visual information and depression. This is followed by a review of related work on multimodal fusion techniques that have been investigated for depression detection systems.

7.2.1 The Relationship between Depression and Language

As discussed in Chapter 2, a significant amount of research has investigated the relationship between prosodic, articulatory, and acoustic features of speech and clinical ratings of depression (Cummins et al., 2015a). In patients with depression, several changes in speech and voice have been noted, including changes in prosody (Blanken et al., 1993), speaking rate (Cannizzaro et al., 2004; Stassen et al., 1998), speech pauses (Alpert et al., 2001), and voice quality (Scherer et al., 2013a).

In addition to voice and speech-based markers, researchers have also provided empirical support for the existence of a relationship between depression and syntax. As discussed in Chapter 2, Section 2.2, depressed individuals exhibit many syntactic patterns including an increased use of first person singular pronouns (Rude et al., 2004) and a decreased use of complex syntactic constructions, such as adverbial clauses (Zinken et al., 2010). The relationship between syntax and depression motivates our syntax-informed fusion approach.

7.2.2 The Relationship between Depression and Facial Activity

Similar to the extensive theoretical and empirical work on the relationship between language and depression, there also exists a body of research on the relationship between depression and facial activity. As discussed in Chapter 2, Section 2.1.3, depression affects individuals' facial expressions, including noted decreases in expressivity, eyebrow movements, and smiling (Cummins et al., 2015a).

In addition, there also exists an interesting relationship between video and audio, e.g. the *McGurk effect*. McGurk and MacDonald (1976) were the first to report a previously unrecognized influence of vision upon speech perception. In their study, they showed participants a video of a young woman speaking, where she repeated utterances of the syllable [ba] which had been dubbed on to lip movements for the syllable [ga]. Participants reported hearing [da]. Then with the reverse dubbing process, a majority reported hearing [bagba] or [gaba]. However, when participants listened to only the sound of the video or when they watched the unprocessed video, they reported the syllables accurately as repetitions of [ba] or [ga]. These findings had important implications for the understanding of speech perception, specifically that visual information a person gets from seeing a person speak changes the way they hear the sound.

These interesting relationships —between the face and voice as well as facial expressions and depression —motivate our video-informed fusion approach.

7.2.3 Existing Fusion Approaches

In recent years, researchers have begun to investigate multimodal features for depression detection systems (Morales et al., 2017b). However, it is a fairly new research interest and as a result only a few studies have compared techniques for fusing features from different modalities (Alghowinem et al., 2015). In the few studies that have investigated fusion techniques, the canonical fusion techniques have been considered, including early, late, and hybrid fusion. In the **early fusion** approach, features are integrated immediately after they are generated through simple concatenation of feature vectors. In the **late fusion** approach integration occurs after each of the modalities have made a decision. In the **hybrid fusion** approach outputs from early fusion and individual unimodal predictors are combined (Baltrusaitis et al., 2017).

Researchers have found early fusion, although simple, to be a successful technique to combine modalities for depression, noting improvements over unimodal systems (Alghowinem et al., 2015; Morales and Levitan, 2016b; Morales et al., 2017b; Scherer et al., 2013c). However, a drawback of the early fusion approach is the high dimensionality of the combined feature vector. Given that drawback, Joshi et al. (2013a) considered early fusion as well as early fusion followed by Principal Component Analysis (PCA), where 98% of the variance was kept. They found that training a depression detection model on this reduced dimensionality feature set led to improved performance of the system over simple early fusion.

Researchers have also investigated late and hybrid fusion. In Alghowinem et al. (2015) a hybrid fusion approach was investigated, which involved concatenating results from individual modalities to the the early fusion feature vector. A majority voting method was used. They evaluated how hybrid fusion and early fusion approaches compare to unimodal approaches. They found that in most cases their early and hybrid fusion models outperformed the unimodal models. Moreover, hybrid fusion models tended to outperform early fusion. Late fusion approaches have also been investigated by some (Joshi et al., 2013a; Meng et al., 2013). For example, Meng et al. (2013) used a late fusion approach that trained a separate model from each modality and combined decisions using the weighted sum rule. They found that combining visual and vocal features at the decision level resulted in further system improvement for depression detection.

Although, in this work, we focus on fusion approaches for depression detection, there exist various studies investigating fusion for other machine learning tasks. Researchers have also proposed new approaches to fusion which differ from the canonical approaches. In particular, deep learning approaches to fusion appear to be particularly promising. For example, Mendels et al. (2017) presented a single hybrid deep model with both acoustic and lexical features trained jointly and found that this approach to fusion achieved state-of-the-art results for deception detection. However, deep learning is not currently a good approach for depression detection, since labeled corpora are not very large and interpretable models are important.

7.3 Dataset

In this work, we use the Distress Analysis Interview Corpus-Wizard of Oz (DAIC-WOZ; Gratch et al., 2014). The corpus is multimodal (video, audio, and transcripts) and is comprised of video interviews between participants and an animated virtual interviewer called Ellie (Figure 7.1), which is controlled by a human interviewer in another room.

Interview participants were drawn from the Greater Los Angeles metropolitan area and included two distinct populations: (1) the general public and (2) veterans of the U.S. armed forces. Participants were coded for depression, Posttraumatic Stress Disorder (PTSD), and anxiety based on accepted psychiatric questionnaires. All participants were fluent English speakers and all interviews were conducted in English. The DAIC-WOZ interviews ranged from 5 to 20 minutes.

The interview started with neutral questions, which were designed to build rapport and make the participant comfortable. The interview then progressed into more targeted questions about symptoms and events related to depression and PTSD. Lastly, the interview

Figure 7.1: Ellie the virtual human interviewer.

ended with a ‘cool-down’ phase, which ensured that participants would not leave the interview in a distressed state. In Table 7.1, we show the contrast between the stages of the interview. Questions from the rapport phase elicit more neutral responses whereas questions from the targeted phase, asked in the second half of the interview, target specific clinical symptoms and therefore elicit more affective responses.

Table 7.1: Description of interview question types.

Context	Question
Rapport phase	How are you doing today?
Rapport phase	Where are you from originally?
Rapport phase	What’d you study at school?
Targeted phase	How are you at controlling your temper?
Targeted phase	Is there anything you regret?
Targeted phase	How easy is it for you to get a good nights sleep?

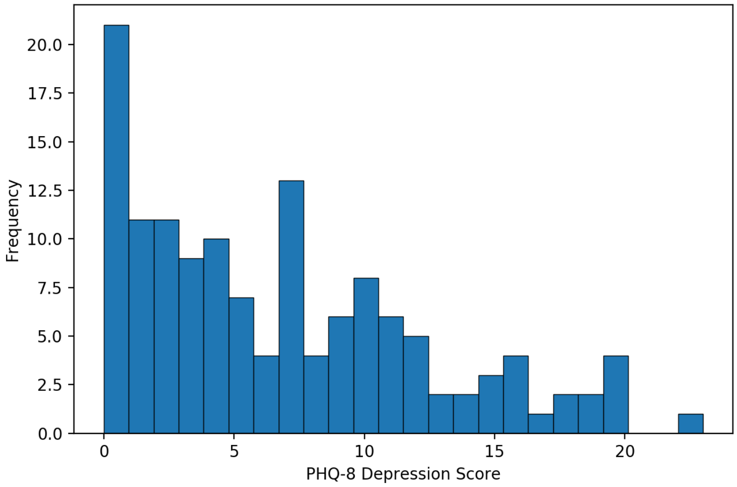


Figure 7.2: Distribution of PHQ-8 depression scores for the subset of the DAIC-WOZ corpus.

The depression label provided includes a PHQ–8[1] score (scale from 0 to 24) as well as a binary depression class label, i.e., score >= 10. As is shown in Figure 7.2 depression scores are skewed. We work with the training and dev splits, which represent interviews from 142 participants.

7.4 Features

In this work we use the OpenMM[2] pipeline to extract multimodal features (Morales et al., 2017b), which uses Covarep (Degottex et al., 2014) and Parsey McParseface (Andor et al., 2016) to extract voice and syntax features.

7.4.1 Voice

In order to extract features from the voice, OpenMM employs Covarep (Degottex et al., 2014). The audio features extracted include prosodic, voice quality, and spectral features.

[1] `http://patienteducation.stanford.edu/research/phq.pdf`

[2] `https://github.com/michellemorales/OpenMM`

Prosodic features include Fundamental frequency (F_0) and voicing boundaries (VUV). Covarep voice quality 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 (peakslope), and shape parameter of the Liljencrants-Fant model of the glottal pulse dynamics (Rd). Spectral features include Mel cepstral coefficients (MCEP0-24), harmonic model and phase distortion mean (HMPDM0-24) and deviations (HMPDD0-12). Lastly, Covarep includes a creak feature which is derived through a creaky voice detection algorithm.

7.4.2 Syntax

In order to generate syntactic features OpenMM employs Google's state-of-the-art pretrained tagger: Parsey McParseface (Andor et al., 2016). For each sentence S, the tagger outputs POS tags. In this work, we make use of 17 POS tags, which are outlined in Table 7.2.

7.4.3 Visual

The visual features we consider are AUs, which were extracted from the DAIC-WOZ corpus as part of the baseline system for the AVEC 2017 challenge (Ringeval et al., 2017). As reviewed in Chapter 3, Section 3.4.1, AUs represent the fundamental actions of individual muscles or groups of muscles. It is a commonly used tool and has become standard to systematically categorize physical expressions, which has proven very useful for psychologists. A detailed list of the facial AUs we consider are given in Table 7.3. Each AU receives a presence score, between -5 and 5, which measures how present that feature is for a given frame of video.

Table 7.2: Description of POS tags.

POS Tag	Description
ADJ	Adjectives
ADV	Adverbs
ADP	Adpositions
AUX	Auxiliaries
CONJ	Conjunctions
DET	Determiners
INTJ	Interjections
NOUN	Nouns
NUM	Cardinal numbers
PPRON	Proper nouns
PRON	Pronouns
PRT	Particles or other functions words
PUNCT	Punctuation
SCONJ	Subordinating conjunctions
SYM	Symbols
VERB	Verbs
X	Other

Table 7.3: Description of facial AUs.

Action Unit	Description
1	Inner brow raise
2	Outer brow raise
4	Brow lowerer
5	Upper lid raiser
6	Check raiser
7	Lid tightener
9	Nose wrinkler
10	Upper lip raiser
12	Lip corner puller
14	Dimpler
15	Lip corner depressor
17	Chin raiser
18	Lip puckerer
20	Lip strecher
23	Lip tightener
24	Lip pressor
25	Lips part
26	Jaw drop
28	Lip suck
43	Eyes closed

7.5 Fusion Approaches

7.5.1 Early Fusion

In our early fusion approach, features are extracted from each modality and then concatenated to generate a single feature vector. Visual and acoustic features are extracted at the frame level while POS tags are extracted at the sentence level. Therefore, the modalities do not align automatically. In order to handle these differences, we first compute statistics (mean, median, standard deviation, maximum, and minimum) across frames/sentences. This results in 370 acoustic features (74 acoustic features × 5 statistical functionals), 100 visual features (20 visual × 5 statistical functionals), and 85 syntactic features (17 syntactic features × 5 statistical functionals). We then fuse the feature vectors to achieve one multimodal feature vector, $features_{early}$.

$$features_{early} = \begin{array}{ccc} \text{Acoustic} & & \text{Syntax} \\ \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_i \end{bmatrix} & + & \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_i \end{bmatrix} \end{array}$$

7.5.2 Informed Early Fusion

Syntax-informed Early Fusion

We compare early fusion to our proposed approach. Our approach leverages syntactic information to target more informative aspects of the speech signal. Given the relationship between depression and syntax, we hypothesize that this approach will help lead to improvements in system performance. First, we align the audio file and transcript file. In order to perform alignment, we use the tool *gentle*[3], which is a forced-aligner built on Kaldi. We then tag each sentence and retrieve the timestamp information for each POS tag. For each POS tag span we extract acoustic features for that time span.

$$features_{mm} = \begin{array}{cc} & \begin{array}{ccc} y_0 & \dots & y_i \end{array} \\ \begin{array}{c} x_0 \\ x_1 \\ \vdots \\ x_i \end{array} & \begin{pmatrix} \bar{x}_0 & \dots & \bar{x}_0 \\ \bar{x}_1 & \dots & \bar{x}_1 \\ \vdots & \vdots & \vdots \\ \bar{x}_i & \dots & \bar{x}_i \end{pmatrix} \end{array}$$

In other words, we are specifically extracting features at the POS level and we are continuously updating our audio features each time we come across a POS tag. For example, each time we see a VERB we use its timestamp information to extract mean F_0 from that specific window and we do this continuously, updating our F_0 value every time we come across a VERB. In the end, we have a mean F_0 value across all VERBs, ADJs, NOUNs, etc., as shown in $features_{mm}$. This representation is different from early fusion in that it conditions the audio features on POS information, providing a representation that does not simply add features from each modality, but instead aims to jointly represent them.

Video-informed Early Fusion

In order to test the robustness of our novel fusion approach —*informed early fusion* —we perform additional experiments using other modalities. The relationship between a person's facial behavior and speech production, motivates our *video-informed fusion* approach. Similar to our syntax-informed approach, where we target POS tags' time frames to identify more informative aspects of the speech signal, we also target aspects of the speech signal using visual information. We hypothesize that targeting informative aspects of the speech

[3] `https://github.com/lowerquality/gentle`

signal using visual cues will help boost system performance when compared with a simple early fusion system.

Similar to syntax-informed fusion, this representation conditions the audio features on AU information. For each frame of video, we identify the AU with the highest presence (value between -5 and 5). Therefore, we assume only one AU can occur per frame. For the AU with the highest presence, we extract acoustic features across that span of time. For each AU, we then aggregate its acoustic features across the entire video. In the end, we have a mean value for each acoustic feature across all AUs.

7.5.3 Late Fusion

We explore two types of late fusion approaches: (1) voting and (2) ensemble. In our voting approach, we train separate classification models for each modality. Each unimodal system makes a classification prediction, depressed or not depressed. We then take the majority vote as our ultimate prediction. We also consider an ensemble approach. In our ensemble late fusion approach, we again train separate classification models for each modality. The models' predictions are then used as features to train a new classification system. The predictions from the newly trained classification system are then used as the final prediction.

7.5.4 Hybrid Fusion

In our hybrid fusion approach, outputs from early fusion and individual unimodal predictors are combined. Therefore, we train separate classification models for each modality. We then take the predictions from each unimodal system and concatenate it with the early fused feature vectors. These new feature vectors (early fusion + unimodal predictors) are then used to train a new model to make the ultimate prediction. We evaluate the hybrid fusion approach with both the early and the *informed fusion* approaches.

7.6 Results

7.6.1 Binary Classification Experiments

In order to evaluate our approach, we conduct a series of participant-level binary classification experiments. We train both unimodal and multimodal models. Our *early + syntax-informed fusion* model combines both the early fusion and syntax-informed fusion feature sets, by early fusion, i.e. simple concatenation. Using scikit-learn[4] we train a Support Vector Machine (SVM) for classification, (linear kernel, C = 0.1). We conduct 5-fold cross-validation on 136 participant interviews (depressed = 26, non-depressed = 110). During cross-validation, each fold is speaker independent and drawn at random. Given the skewness of the dataset, we set the SVM model's class weight parameter to 'balanced', which automatically adjusts the weights of the model inversely proportional to the class frequencies in the data, helping adjust for the class imbalance. Given the possibility of sparse feature values and the differences in dimensionality across feature sets, we also perform feature selection. We use scikit-learn's *Select K-Best* feature selection approach, which computes the ANOVA F-value across features and identifies the K most significant features. We set K to 20 and evaluate each feature set's best set. We report our findings in Table 7.4. We report precision, recall, and F1-score for the depressed class. We choose to report these values instead of the average values across both classes because the depressed class label is the harder class to detect. As a result, the non-depressed class usually reports very high scores which tend to inflate the average score. If we can increase the performance of the depressed class, it can be assumed that the overall performance will go up as a result.

We find that the novel syntax-informed fusion approach performs best, with an F1-score of 0.49. We believe this approach is able to leverage syntactic information to target more informative aspects of the speech signal resulting in higher performing models. By

[4] `http://scikit-learn.org/`

Table 7.4: Results for 5-fold cross-validation using SVM.

Modality	Fusion Type	Precision	Recall	F1-score
A	–	0.34	0.70	0.45
S	–	0.21	0.96	0.35
V	–	0.16	0.52	0.25
A + S	E	0.34	0.70	0.45
A + S	I	0.40	0.69	**0.49**
A + S	E + I	0.36	0.62	0.44
A + V	E	0.37	0.70	0.48
A + V	I	0.36	0.77	**0.49**
A + V	E + I	0.34	0.74	0.46

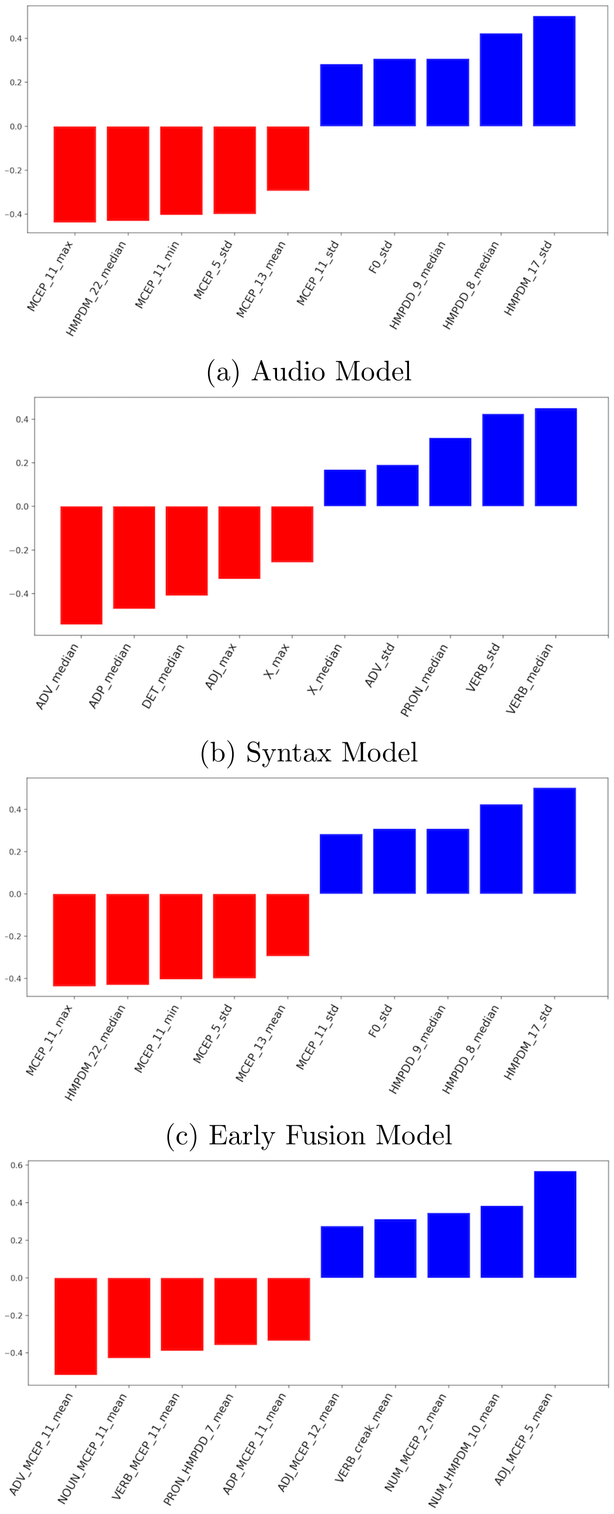
Note. Results reported for the audio (A), syntax (S), video (V), and fusion (A + S) approaches. Fusion types include early (E), syntax-informed (I), and both (E + I).

conditioning acoustic models on syntactic information this approach combines information from both modalities in a way a human clinician might. Syntax-informed fusion substantially outperforms early fusion in precision and F1-score. In recall, performance is similar for both approaches. In addition, the syntax-informed method surfaces novel multimodal features. For example, creak is not a useful feature in the early fusion or the acoustic model. However, when we consider verb creak we find it extremely useful. This is demonstrated in Figure 7.3. To better understand each model, we inspect the coefficient weights of the SVM models. Using the weight coefficients from the models, we plot the top 5 most important features by class in Figure 7.3. The absolute size of the coefficients in relation to each other can be used to determine feature importance for the depression detection task.

If we consider the audio and early fusion models in Figure 7.3a and Figure 7.3c, we find that both models weight the same features highly. Although the early fusion model also includes the set of syntax features, it still prefers the same five features as the audio-only model. Since early fusion is simply concatenating the audio and syntax feature vectors it is understandable to find similar features performing well. These results show the promise of these specific audio features, which include spectral and prosodic (F_0) features. These results support previous work that showed spectral and prosodic features were useful for detecting depression (Cummins et al., 2015a).

However, these findings also highlight the limitation of early fusion. The intention behind early fusion is to have access to multiple modalities that observe the same phenomenon to allow for more robust predictions, allowing for complementary information from each modality. Something not visible in individual modalities may appear when using multiple modalities. However, in early fusion, we can not guarantee that information from both modalities is considered. For example, if we inspect the feature set for early fusion we find that no syntax features appear; this could be attributed to the strength of the audio features as well as the difference in dimensionality size between the audio and syntax sets; the audio feature set is almost 5 times larger than the syntax set.

The syntax-informed fusion model is promising because it does not possess the same limitation as early fusion; with syntax-informed fusion we can guarantee that information from both modalities is considered. This could also be considered a drawback of syntax-informed fusion, in circumstances where one would like to be agnostic regarding the value of each modality. However, in a task for which multiple modalities are known to be important and interconnected, such as depression detection, it is valuable to represent them jointly. Figure 7.3d demonstrates that syntax-informed fusion is able to capture important information from both modalities. We find the best features used to distinguish between classes



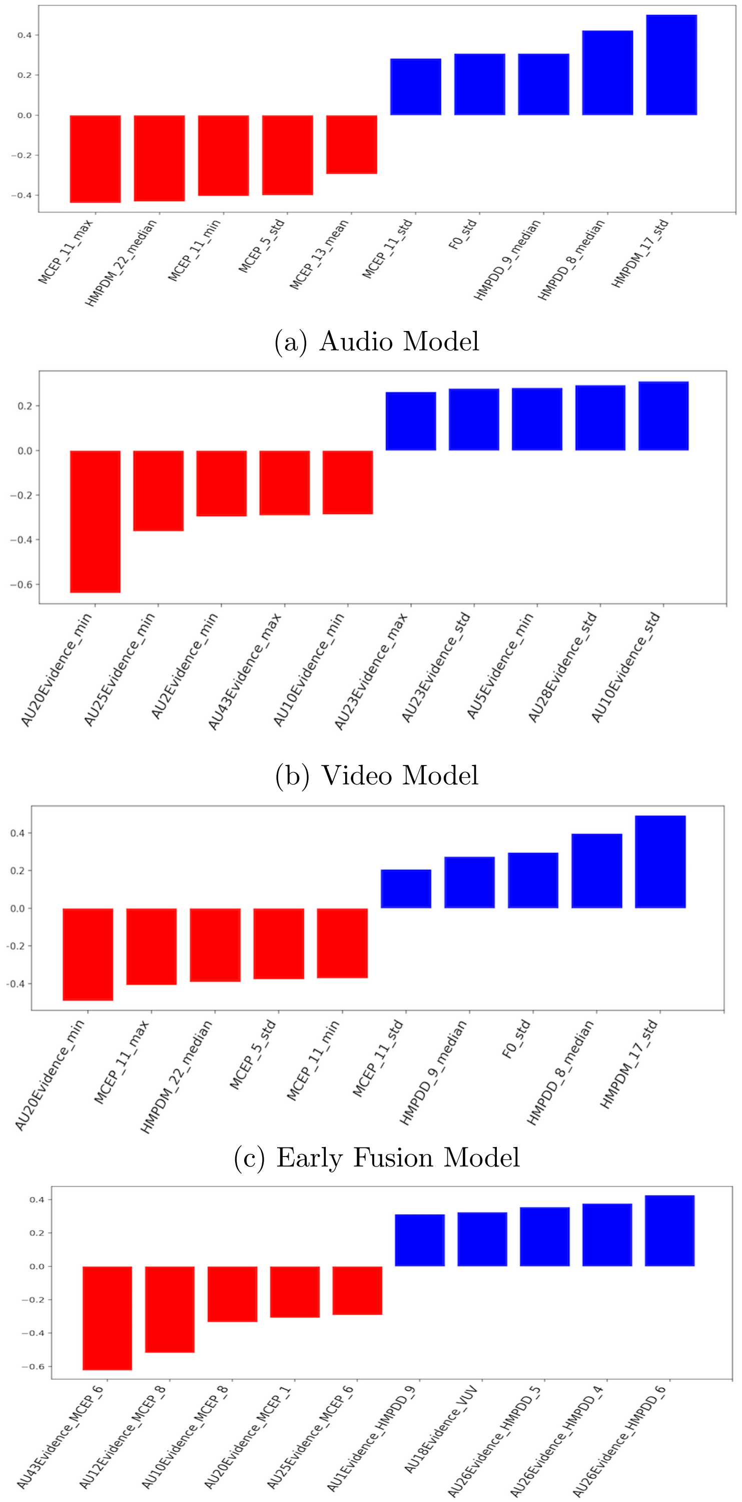
(a) Audio Model

(b) Syntax Model

(c) Early Fusion Model

(d) Syntax-informed Fusion Model

Figure 7.3: Illustration of linear kernel SVM's coefficient weights by class. Blue bars represent the positive or depressed class. Red bars represent negative or healthy class.



(a) Audio Model

(b) Video Model

(c) Early Fusion Model

(d) Video-informed Fusion Model

Figure 7.4: Illustration of linear kernel SVM's coefficient weights by class. Blue bars represent the positive or depressed class. Red bars represent negative or healthy class.

are spectral features that span the production of pronouns, verbs, and adverbials. In other words, the best syntax-informed features represent a fused multimodal representation of the best features from each unimodal domain.

We also find further support of the relationship between depression and syntax. From the syntax-only model, we find pronouns (PRON) to be useful in identifying the depressed class, which supports previous findings that pronoun use can help identify depression (Rude et al., 2004). In addition, we find the POS tag category X (other) to be useful in distinguishing between classes. After manually inspecting the transcripts, we find the X POS tag is often assigned to filler words such as *uh, um, mm.* These results suggest filler words can be helpful in identifying depression. Lastly, we find adverbials (ADV) to be useful in distinguishing between classes. These results are especially interesting because Zinken et al. (2010) argued that adverbial clauses could help predict the improvement of depression symptoms. To the best of our knowledge, these results are the first to show support that adverbial clauses could also help predict depression.

We find similar results for video-informed fusion. Video-informed fusion outperforms early fusion in recall and F1-score. Similar to syntax-informed fusion we find that video-informed fused features are able to jointly capture the most informative features from each individual modality. For example, we find the best performing acoustic features and AUs from the unimodal systems to appear together in the video-informed system.

7.6.2 Fusion Experiments

In addition to evaluating how well our novel approach compares to early fusion, we also evaluate other types of fusion such as late and hybrid fusion. These series of experiments follow the same configuration as our first series of experiments: 5 fold cross-validation using SVM (linear kernel, C = 0.1, class weights balanced). We evaluate each method of fusion —early, informed, late (vote/ensemble), and hybrid (early/informed) —and report our results in

Table 7.5.

As mentioned previously, in regards to early fusion methods, the *informed fusion* approaches outperform simple early fusion. When we compare the syntax and video-informed fusion techniques with other approaches, such as late and hybrid fusion, we do not find differences between the systems. When we evaluate systems that use all three modalities (A + S + V), we find a late ensemble approach performs best. We also find that late fusion techniques which rely on voting perform the worst. We believe these results can be attributed to the low performing unimodal video system, as demonstrated in Table 7.4. This finding highlights a weakness of the late fusion (voting) approach. Since it weighs the prediction from each system equally, this can lead to poor performance when one of the unimodal systems is weak.

7.6.3 Regression Experiments

Our last series of experiments, evaluates how well these systems fare in predicting depression PHQ-8 score. We build a Support Vector Regression model in scikit-learn using its libsvm implementation. For the regression models, we report MAE. In addition to considering MAE we also report R^2 and adjusted R^2, the coefficient of determination. R^2 provides a measure of how well future samples are likely to be predicted by the model, i.e. how well the model captures the variance of the data. For R^2, the best possible score is 1.0. Results are shown in Table 7.6.

In terms of error, we find the early fusion system (A + S + V) performs best for the regression task. However, if we consider R^2, we find that all models receive low scores which suggests that the models do not fit the data well and are not capturing the variance of the data. We considered the skewness of the data set as a possible factor leading to the poor performance. Therefore, we upsampled data points above a PHQ-8 score of 10. This created a more balanced distribution of scores. However, we found this systematically led to poorer

Table 7.5: Results for fusion experiments using SVM.

Modality/Features	Fusion Type	Precision	Recall	F1-score
A + S	Early	0.34	0.70	0.45
A + S	Informed	0.40	0.69	0.49
A + S	Late - ensemble	0.36	0.78	0.49
A + S	Hybrid - informed	0.36	0.78	0.49
A + S	Hybrid - early	0.34	0.74	0.46
A + V	Early	0.37	0.70	0.48
A + V	Informed	0.36	0.77	0.49
A + V	Late - ensemble	0.36	0.78	0.49
A + V	Hybrid - informed	0.36	0.78	0.49
A + V	Hybrid - early	0.50	0.74	0.35
A + S + V	Early	0.37	0.70	0.48
A + S + V	Late - vote	0.50	0.17	0.25
A + S + V	Late - ensemble	0.36	0.78	0.49

Note. Results for fusion approaches including features from audio (A), syntax (S), and video (V).

Table 7.6: Results for 5-fold cross-validation using SVM regression.

Modality	Fusion Type	MAE	R^2	Adjusted R^2
A	–	4.93	0.03	-0.26
S	–	5.03	-0.03	-0.31
V	–	5.01	-0.01	-0.29
A + S	Early	4.93	0.03	-0.26
A + S	Informed	4.90	0.02	-0.24
A + V	Early	4.93	0.03	-0.25
A + V	Informed	5.00	0.01	-0.27
A + S + V	Early	2.43	0.02	-0.04

Note. Results reported for the audio (A), syntax (S), and video (V) models. Number of participants $N = 136$. Each modality represents the best 20 features from that set.

performance. The initial focus of this work was to build binary depression detection systems. These results demonstrate that a system meant to predict depression class will not necessarily perform well in a regression task. However, a system that predicts depression score well, will arguably do well in predicting depression class. This notion should be considered when outlining the framework for a detection system. In addition, these findings also highlight the misleading nature of error metrics. Currently, state-of-the-art systems for depression detection tend to only report MAE or RMSE. Therefore, it is difficult to determine how well our systems compare when measured by R^2. For example, the AVEC 2017 depression challenge, which can be viewed as setting the standard for depression detection system performance, only report MAE and RMSE (Ringeval et al., 2017). In 2017, they reported an MAE of 5.52 for their best system using the train/dev splits. The winner of the challenge reported a MAE of 3.96 on the test set. If we consider our performance as compared to these challenge baselines, we find our system performs well, matching or outperforming the state-of-the-art. Future work, will consider the regression task further.

7.7 Discussion

In this chapter, we present a novel approach to early fusion: *informed fusion*. The syntax-informed fusion approach is able to leverage syntactic information to target more informative aspects of the speech signal. We find that syntax-informed early fusion approach outperforms early fusion. Given some of the limitations to early fusion, we believe syntax-informed early fusion is a promising alternative dependent on the classification task. In addition, we evaluate this approach's robustness by evaluating the technique with other modalities. Specifically, we evaluate video-informed fusion and confirm our findings that *informed fusion* outperforms early fusion. We also confirm previous findings that spectral features and prosodic features are useful in identifying depression. In addition, we present further support for the relationship between syntax and depression. Specifically we find pronouns, adverbials, and fillers to be useful in identifying individuals with depression. Lastly, we perform an in-depth investigation of fusion techniques and find that *informed*, late, and hybrid approaches perform comparably. To the best of our knowledge, this work represents the most comprehensive empirical study of fusion techniques for multimodal (audio, video, and text) depression detection. However, this analysis is conducted on one dataset. Future work will consider extending this study to include many of the publicly-available existing datasets mentioned in Chapter 3.

Chapter 8

Conclusions

Depression is a serious illness that affects a large portion of the world's population. Given the large effect it has on society, it is evident that depression is a serious health issue. Therefore, it is imperative that the research community continues to investigate how best to diagnose and treat depression. This thesis evaluates, at length, how technology may aid in assessing depression. More specifically, this thesis focuses on how depression detection systems may serve to diagnose depression, presenting a broad study of features and fusion techniques.

8.1 Limitations and Future Work

We have discussed limitations and possibilities for future work throughout the thesis. In this section, we summarize these limitations and describe several directions that arise from the thesis as a whole.

- In this work, we only use datasets that include self-report measures for depression. While the PHQ-8 and BDI-II surveys are well-validated measures of depression, the relationships we find between features and depression labels are not necessarily the result of depression and could be attributed to other factors such as general distress

(Calvo et al., 2017b). Future work should consider using self-report measures as well as possible alternatives, such as a structured clinical interview or a rating scale.

- Individuals suffering from depression represent a heterogenous group. As a result, certain depression diagnostic measures were developed specifically for a particular population of depressed individuals (Gotlib and Hammen, 2014). Some variables that can affect diagnostic appropriateness include age, gender, comorbidity, and cultural differences. For example, depression affects about twice as many women as men; the National Comorbidity Survey found a lifetime prevalence for major depressive disorder of 21.3% in women and 12.7% in men (Kessler et al., 1993). In addition, 50% of Parkinson's disease patients may experience depression, 50-75% of eating disorder patients experience depression, and 25% of cancer patients experience depression. Comorbidity has the potential to limit the validity of depression measures that do not take these disorders into account. As a result, specific assessment instruments have been developed, such as the Calgary Depression Scale for Schizophrenia, which was created to address the fact that current depression diagnostics did not accurately represent depressive symptoms or syndromes in persons with schizophrenia. In regards to cultural differences, it is not enough to ensure that a self-report measure has been correctly translated into another language. It is also important to demonstrate that the survey correctly addresses the constructs that have meaning within a particular culture. Studies have shown (Gotlib and Hammen, 2014) that although similarities exist in the expression of depression across various cultures, differences do exist. Although we touch upon these issues in Chapter 6, these factors should be investigated thoroughly any time a system is built. Future work should consider recruiting clinical populations, varying age groups, and culturally distinct groups to better understand how these factors may affect system performance. In addition, further research should be closely integrated with ongoing psychological research regarding the heterogeneity of depression.

- While we do have participants who report moderate and severe levels of depression, it represents a very small sample. It is possible that the characteristics this group exhibits are not representative of typical trends seen in people with depression. Future work could explicitly recruit more participants from the moderate and severe levels to achieve a more representative population.

- The labels we use represent a snapshot of a person's mental state. It represents the individual's current score of exhibited depression symptoms. It does not represent continuous data of a person's state. Therefore, it is difficult to gauge how a system would fare over time. Future work should consider a longitudinal study where depression detection system performance is evaluated over time.

- Technical limitations:

  - Many of the systems presented in this thesis incorporate an ASR system. Although ASR systems have improved dramatically in recent years, they are still sensitive and require clean audio to perform well. The datasets we use were recorded under ideal conditions, presenting clear speech. Future work should evaluate how well these systems can perform under less than ideal recording conditions.

  - Our tool, OpenMM, has been publicly shared on GitHub. However, it is still in a codebase form. In order to facilitate the use of this tool for non-computational researchers, future work should include the release of a graphical user interface that can be used to easily run the tool.

  - The ASR systems used by OpenMM include Watson and Google's speech-to-text systems. In both cases, an API is called to transform the audio files into transcripts. Given the sensitive nature of depression datasets, future work should include a more private alternative such as Kaldi: a local ASR system (Povey et al., 2011).

8.2 Societal Impact

Given the sensitive nature of building technologies for psychological assessment, it is important to discuss the societal impacts of depression detection systems. As with any technology or tool, there is always risk of misuse and therefore it is important to discuss general ethical considerations with pursuing this line of research. It is especially important to define and outline appropriate use of these systems. Mental health professionals should view language technology for depression detection as a mechanism to complement current diagnoses by giving them access to a novel and rich non-intrusive data source. It is understandable that mental health professionals as well as the general population may be uncomfortable with the possibility that technologies might have to predict psychological states, especially when relatively accurate predictions can be made. To be clear, these systems are not proposed as standalone diagnostic tools that could replace current approaches to diagnosing mental health issues, but instead proposed as part of a broader awareness, detection, and support system. These technologies provide numerous advantages, including large-scale and remote assessment, which in turn could help a broader population. These methods could also provide a lower cost complement to traditional depression assessments. In addition, these tools could help health professionals manage current patients more efficiently, allowing clinicians to monitor their patients continuously. Determining how machines should augment and assist in diagnosis is a complicated issue. However, there exists evidence that mechanical prediction (statistical, algorithmic, etc.) is typically as accurate or more accurate than clinical prediction (Grove et al., 2000). Moreover, mechanical predictions do not require an expert judgment and are completely reproducible. Although there are general ethical considerations, it is important to highlight the potential of mental health assessment tools to enhance the quality of life for society.

It is also important to understand that these detection systems, and similar automated technologies, are inherently interdisciplinary. Therefore, it is necessary that all disciplines work together to build, evaluate, and understand these systems. Each discipline contributes a necessary piece of understanding and without collaboration between all disciplines, it will be difficult to create a system that can ultimately be used in practice. Therefore, collaboration is extremely crucial to future work. However, in circumstances where collaboration between fields is difficult, it is possible to help mitigate this issue by building systems and considering results from an interdisciplinary perspective. This can be done by building hypothesis driven interpretable systems motivated from theory, such as linguistics or psychology. In doing so, this will allow for a better understanding of performance, allowing researchers to make connections between theory and empirical work. In addition, it is important to adopt evaluation approaches from each field. Presenting a range of evaluation metrics will help standardize evaluation and allow for systematic comparisons across fields. Therefore, during each stage of research —understanding related work, creating methodology, and performing evaluation —an interdisciplinary perspective should be adopted.

Lastly, in order for the research community to progress together, researchers should begin to follow the best practices (Stodden and Miguez, 2013) that establish communication standards, which will help disseminate reproducible research, facilitate innovation by enabling data and code re-use, and enable broader communication of the output of computational research. Without the data and code that underlie scientific discoveries, it is all but impossible to verify published findings. We urge researchers to focus on reproducible research, through the dissemination, availability, and accessibility of data and code.

8.3 Contributions

This thesis presents the following novel contributions:

- An in-depth multidisciplinary survey of theoretical and empirical depression research spanning psychology, computer science, and linguistics.

- An analysis and survey of depression detection systems across research fields and modalities. To the best of our knowledge, this survey is the first comprehensive review of depression detection systems that spans all modalities. The review discusses existing methodologies highlighting the most promising approaches to this task. This survey helps contribute an improved understanding of current research.

- A comparative study of text-based and speech-based depression detection systems. To our knowledge, this work presented the first ever automated speech to text-based system for depression detection. We find that a multimodal system derived from both unimodal systems performs best. This analysis also provides an improved understanding of the potential of ASR in a depression detection system. As a consequence, we stress the value of incorporating an ASR into the depression detection system pipeline. There exist many stages to the speech production process and researchers should draw from as many stages as possible, including the phonetic, semantic, and syntactic level. We also find strong correlations between word use and depression level. Findings from this investigation of linguistic features for depression provide support for previous findings that found pronoun use and negatively-valenced words to be correlated with depression. We also provide new evidence that a relationship exists between language related to work/sleep and depression level.

- An open-source multimodal feature extraction tool: **OpenMM**. To the best of our knowledge, we present the first publicly available tool for multimodal feature extraction. OpenMM provides researchers with a simple way of extracting multimodal features and consequently a richer and more robust feature representation for machine learning tasks.

- An evaluation of OpenMM's multimodal (visual, acoustic, and linguistic) features. We demonstrate OpenMM's feature set's robustness, finding it matches state-of-the-art performance in three machine learning tasks: depression detection, deception detection, and sentiment detection.

- A systematic evaluation of unimodal and multimodal depression detection systems. We provide a comparative analyses of various features for depression detection on several datasets. We find in almost all cases, across languages, a multimodal system outperforms a unimodal system.

- We present the first thorough investigation of fusion techniques for multimodal depression detection systems

- We present a novel fusion technique: *informed fusion*. We evaluate our proposed approach against existing techniques and find our approach achieves the best performance for this task.

- We present an analysis of confounding factors for depression. We discuss approaches to handling confounding factors and present a novel technique to mitigate such factors, which uses a multi-step approach that performs unsupervised clustering prior to depression classification.

The contributions outlined above, serve to address the primary research goals of this thesis, which were given in Chapter 1:

1. Bridge the disconnect that exists between the depression detection research fields

2. Investigate fusion techniques for multimodal depression detection systems

3. Develop and release an open-source tool

The primary goal of this work was to help bridge the disconnect that exists between the depression detection research fields. This research goal was the central aim of this thesis and each subsequent research goal served to contribute to this primary objective. This goal was of primary importance because, in order to build an accurate depression detection system, collaboration between research fields is necessary. Each research field provides one important piece of understanding, that is necessary to build a comprehensive detection system that can represent and measure an individual's entire behavior. Therefore, to truly build an all-inclusive and accurate system collaboration and connection between fields is necessary. This thesis provides the research community with a comprehensive resource, which we hope will serve as a bridge to provide improved understanding of existing work and help shape future directions.

In this thesis we present an in-depth investigation of features and fusion techniques for depression detection systems. In addition, we present OpenMM: a novel tool for multimodal feature extraction. We also present novel techniques for multimodal fusion. The contributions of this work add considerably to our knowledge of depression detection systems and have the potential to improve future systems by incorporating that knowledge into their design, improving assessment performance.

Bibliography

Abel, L., Friedman, L., Jesberger, J. A., Malki, A. E., and Meltzer, H. Y. (1991). Quantitative assessment of smooth pursuit gain and catch-up saccades in schizophrenia and affective disorders. *Biological psychiatry*, 29 11:1063–72.

Alghowinem, S., Goecke, R., Cohn, J. F., Wagner, M., Parker, G., and Breakspear, M. (2015). Cross-cultural detection of depression from nonverbal behaviour. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 1, pages 1–8. IEEE.

Alghowinem, S., Goecke, R., Wagner, M., Epps, J., Gedeon, T., Breakspear, M., and Parker, G. (2013). A comparative study of different classifiers for detecting depression from spontaneous speech. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8022–8026. IEEE.

Alpert, M., Pouget, E. R., and Silva, R. R. (2001). Reflections of depression in acoustic measures of the patients speech. *Journal of Affective Disorders*, 66(1):59–69.

Althoff, T., Clark, K., and Leskovec, J. (2016). Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *arXiv preprint arXiv:1605.04462*.

American Psychological Association (2013). *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub.

Amos, B., Ludwiczuk, B., and Satyanarayanan, M. (2016). Openface: A general-purpose face recognition library with mobile applications. Technical report, CMU-CS-16-118, CMU School of Computer Science.

Anderson-Sprecher, R. (1994). Model comparisons and r2. *The American Statistician*, 48(2):113–117.

Andor, D., Alberti, C., Weiss, D., Severyn, A., Presta, A., Ganchev, K., Petrov, S., and Collins, M. (2016). Globally normalized transition-based neural networks. *arXiv preprint arXiv:1603.06042*.

Anxiety and Depression Association of America (2015). Facts & Statistics, Anxiety and Depression, Association of America, ADAA.

Balsters, M Emiel Krahmer, M. S. and Vingerhoets, A. (2012). Verbal and nonverbal correlates for depression: a review.

Baltrusaitis, T., Ahuja, C., and Morency, L.-P. (2017). Multimodal machine learning: A survey and taxonomy. *CoRR*, abs/1705.09406.

Baltrušaitis, T., Robinson, P., and Morency, L.-P. (2016). Openface: an open source facial behavior analysis toolkit. In *IEEE Winter Conference on Applications of Computer Vision*.

Beck, A. T. (1967). *Depression: Clinical, experimental, and theoretical aspects*. University of Pennsylvania Press.

Beck, A. T., Ward, C., Mendelson, M., et al. (1961). Beck depression inventory (bdi). *Arch Gen Psychiatry*, 4(6):561–571.

Becker, J. T., Boiler, F., Lopez, O. L., Saxton, J., and McGonigle, K. L. (1994). The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archives of Neurology*, 51(6):585–594.

Blanken, G., Dittmann, J., Grimm, H., Marshall, J. C., and Wallesch, C.-W. (1993). *Linguistic disorders and pathologies: an international handbook*, volume 8. Walter de Gruyter.

Bos, E. H., Geerts, E. A. H. M., and Bouhuys, A. L. (2002). Non-verbal interaction involvement as an indicator of prognosis in remitted depressed subjects. *Psychiatry research*, 113 3:269–77.

Brew, C. (2016). Classifying reachout posts with a radial basis function svm. *red*, 14(23):27.

Bucci, W. and Freedman, N. (1981). The language of depression. *Bulletin of the Menninger Clinic*, 45(4):334.

Calvo, R. A., Milne, D. N., Hussain, M. S., and Christensen, H. (2017a). Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*, page 137.

Calvo, R. A., Milne, D. N., Hussain, M. S., and Christensen, H. (2017b). Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*, 23:649–685.

Cannizzaro, M., Harel, B., Reilly, N., Chappell, P., and Snyder, P. J. (2004). Voice acoustical measurement of the severity of major depression. *Brain and cognition*, 56(1):30–35.

Carney, R. M., Freedland, K. E., and Veith, R. C. (2005). Depression, the autonomic nervous system, and coronary heart disease. *Psychosomatic medicine*, 67:S29–S33.

Chai, T. and Draxler, R. R. (2014). Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature. *Geoscientific Model Development*, 7(3):1247–1250.

Chomsky, N. and Halle, M. (1968). The sound pattern of english.

Choudhury, M. D., Gamon, M., Counts, S., and Horvitz, E. (2013). Predicting depression via social media. In *ICWSM*.

Christopher, G. and MacDonald, J. (2005). The impact of clinical depression on working memory. *Cognitive Neuropsychiatry*, 10(5):379–399.

Cohn, J. F., Kruez, T. S., Matthews, I., Yang, Y., Nguyen, M. H., Padilla, M. T., Zhou, F., and La Torre, F. D. (2009). Detecting depression from facial actions and vocal prosody. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pages 1–7. IEEE.

Cooper, W. E. and Paccia-Cooper, J. (1980). *Syntax and speech.* Number 3. Harvard University Press.

Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., and Mitchell, M. (2015). Clpsych 2015 shared task: Depression and ptsd on twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39.

Croarkin, P. E., Levinson, A., and Daskalakis, Z. J. (2011). Evidence for gabaergic inhibitory deficits in major depressive disorder. *Neuroscience and biobehavioral reviews*, 35 3:818–25.

Cummins, N., Epps, J., Breakspear, M., and Goecke, R. (2011). An investigation of depressed speech detection: Features and normalization. In *Interspeech*, pages 2997–3000.

Cummins, N., Epps, J., Sethu, V., and Krajewski, J. (2014). Variability compensation in small data: Oversampled extraction of i-vectors for the classification of depressed speech. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 970–974. IEEE.

Cummins, N., Joshi, J., Dhall, A., Sethu, V., Goecke, R., and Epps, J. (2013). Diagnosis of depression by behavioural signals: a multimodal approach. In *3rd ACM international workshop on Audio/visual emotion challenge Proc.*, pages 11–20. ACM.

Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., and Quatieri, T. F. (2015a). A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71:10–49.

Cummins, N., Sethu, V., Epps, J., Schnieder, S., and Krajewski, J. (2015b). Analysis of acoustic space variability in speech affected by depression. *Speech Communication*, 75:27–49.

Darby, J. and Hollien, H. (1977). Vocal and speech patterns of depressive patients. *Folia Phoniatrica et Logopaedica*, 29(4):279–291.

De Choudhury, M., Counts, S., Horvitz, E. J., and Hoff, A. (2014). Characterizing and predicting postpartum depression from shared facebook data. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 626–638. ACM.

Degottex, G., Kane, J., Drugman, T., Raitio, T., and Scherer, S. (2014). Covarepa collaborative voice analysis repository for speech technologies. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 960–964. IEEE.

D'Mello, S. and Kory, J. (2012). Consistent but modest: a meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 31–38. ACM.

Durkheim, E. (1951). Suicide (g. simpson, trans.).

Ekman, P., Friesen, W. V., and Hager, J. C. (1978). Facial action coding system (facs). *A technique for the measurement of facial action. Consulting, Palo Alto*, 22.

Ellgring, H. and Scherer, K. (1996). Vocal indicators of mood change in depression. *Journal of Nonverbal Behavior*, 20:83–110.

Eyben, F., Weninger, F., Gross, F., and Schuller, B. (2013). Recent developments in opensmile, the munich open-source multimedia feature extractor. In *21st ACM international conference on Multimedia Proc.*, pages 835–838. ACM.

Eyben, F., Wöllmer, M., and Schuller, B. (2009). Openearintroducing the munich open-source emotion and affect recognition toolkit. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pages 1–6. IEEE.

Ferrer, A. S. M. A. L., Kajarekar, S., and Shriberg, C. R. N. S. E. Improving language recognition with multilingual phone recognition and speaker adaptation transforms.

Fraser, K. C., Rudzicz, F., and Hirst, G. (2016). Detecting late-life depression in alzheimers disease through analysis of speech and language. *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology.*

Freud, S. (1901). 1960. the psychopathology of everyday life. *The standard edition of the complete psychological works of Sigmund Freud*, 6.

Freud, S. (1917). 1969. mourning and melancholia. *The standard edition of the complete psychological works of Sigmund Freud*, 14:239–60.

Gaebel, W. and Wolwer, W. (2004). Facial expressivity in the course of schizophrenia and depression. *European Archives of Psychiatry and Clinical Neuroscience*, 254:335–342.

Girard, J. M., Cohn, J. F., Mahoor, M. H., Mavadati, S. M., Hammal, Z., and Rosenwald, D. P. (2014). Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses. *Image and vision computing*, 32(10):641–647.

Gotlib, I. and Hammen, C. (2014). *Handbook of Depression, Third Edition*. Guilford Publications.

Gratch, J., Artstein, R., Lucas, G. M., Stratou, G., Scherer, S., Nazarian, A., Wood, R., Boberg, J., DeVault, D., Marsella, S., et al. (2014). The distress analysis interview corpus of human and computer interviews. In *LREC*, pages 3123–3128.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., and Nelson, C. (2000). Clinical versus mechanical prediction: a meta-analysis. *Psychological assessment*, 12(1):19.

Hall, J. A. and Rosenthal, R. W. (2005). Nonverbal behavior in clinician-patient interaction.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Hall, M. A. (1999). *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato.

Hasler, G., Drevets, W. C., Manji, H. K., and Charney, D. S. (2004). Discovering endophenotypes for major depression. *Neuropsychopharmacology*, 29(10):1765–1781.

Hönig, F., Batliner, A., Nöth, E., Schnieder, S., and Krajewski, J. (2014). Automatic modelling of depressed speech: relevant features and relevance of gender. In *INTERSPEECH*, pages 1248–1252.

Horwitz, R., Quatieri, T. F., Helfer, B. S., Yu, B., Williamson, J. R., and Mundt, J. (2013). On the relative importance of vocal source, system, and prosody in human depression. In *Body Sensor Networks (BSN), 2013 IEEE International Conference on*, pages 1–6. IEEE.

Hu, Y., Wu, D., and Nucci, A. (2013). Fuzzy-clustering-based decision tree approach for large population speaker identification. *Audio, Speech, and Language Processing, IEEE Transactions on*, 21(4):762–774.

Joshi, J., Goecke, R., Alghowinem, S., Dhall, A., Wagner, M., Epps, J., Parker, G., and Breakspear, M. (2013a). Multimodal assistive technologies for depression diagnosis and monitoring. *Journal on Multimodal User Interfaces*, 7(3):217–228.

Joshi, J., Goecke, R., Parker, G., and Breakspear, M. (2013b). Can body expressions contribute to automatic depression analysis? In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–7. IEEE.

Kessler, R. C., McGonagle, K. A., Swartz, M. S., Blazer, D. G., and Nelson, C. B. (1993). Sex and depression in the national comorbidity survey. i: Lifetime prevalence, chronicity and recurrence. *Journal of affective disorders*, 29 2-3:85–96.

Kim, S. M., Wang, Y., Wan, S., and Paris, C. (2016). Data61-csiro systems at the clpsych 2016 shared task. In *CLPsych@HLT-NAACL*.

Kinnunen, T., Kilpeläinen, T., and Fränti, P. (2011). Comparison of clustering algorithms in speaker identification. *dim*, 1:2.

Klakow, D. and Peters, J. (2002). Testing the correlation of word error rate and perplexity. *Speech Communication*, 38(1):19–28.

Kroenke, K., Spitzer, R. L., and Williams, J. B. (2001). The phq-9. *Journal of general internal medicine*, 16(9):606–613.

Lenneberg, E. H., Chomsky, N., and Marx, O. (1967). *Biological foundations of language*, volume 68. Wiley New York.

Lewinsohn, P. M. (1974). A behavioral approach to depression.

Lieberman, H. A. and Meyer, A. R. (2014). Visualizations for mental health topic models. In *Massachusetts Institute of Technology Master's Thesis*.

Lipton, R. B., Levin, S., and Holzman, P. S. (1980). Horizontal and vertical pursuit eye movements, the oculocephalic reflex, and the functional psychoses. *Psychiatry research*, 3 2:193–203.

Littlewort, G., Whitehill, J., Wu, T., Fasel, I., Frank, M., Movellan, J., and Bartlett, M. (2011). The computer expression recognition toolbox (cert). In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 298–305. IEEE.

Low, L.-S. A., Maddage, N. C., Lech, M., Sheeber, L. B., and Allen, N. B. (2011). Detection of clinical depression in adolescents speech during family interactions. *IEEE Transactions on Biomedical Engineering*, 58(3):574–586.

Maddage, N. C., Senaratne, R., Low, L.-S. A., Lech, M., and Allen, N. B. (2009). Video-based detection of the clinical depression in adolescents. *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, 2009:3723–6.

Malmasi, S., Zampieri, M., and Dras, M. (2016). Predicting post severity in mental health forums. *order*, 2:8.

McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264:746–748.

McLeod, S. (2015). Psychological theories of depression.

Mendels, G., Levitan, S. I., Lee, K.-Z., and Hirschberg, J. (2017). Hybrid acoustic-lexical deep learning approach for deception detection.

Meng, H., Huang, D., Wang, H., Yang, H., AI-Shuraifi, M., and Wang, Y. (2013). Depression recognition based on dynamic facial and vocal expression features using partial least square regression. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 21–30. ACM.

Milne, D. N., Pink, G., Hachey, B., and Calvo, R. A. (2016). Clpsych 2016 shared task: Triaging content in online peer-support forums. In *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*, pages 118–127.

Montgomery, S. A. and Asberg, M. (1979). A new depression scale designed to be sensitive to change. *The British journal of psychiatry*, 134(4):382–389.

Moore, E., Clements, M., Peifer, J. W., Weisser, L., et al. (2008). Critical analysis of the impact of glottal features in the classification of clinical depression in speech. *Biomedical Engineering, IEEE Transactions on*, 55(1):96–107.

Morales, M. R. and Levitan, R. (2016a). Mitigating confounding factors in depression detection using an unsupervised clustering approach. In *Proceedings of the 2016 Computing and Mental Health Workshop*.

Morales, M. R. and Levitan, R. (2016b). Speech vs. text: A comparative analysis of features for depression detection systems. In *Spoken Language Technology Workshop (SLT), 2016 IEEE*, pages 136–143. IEEE.

Morales, M. R., Scherer, S., and Levitan, R. (2017a). A Cross-modal Review of Indicators for Depression Detection Systems. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology*, pages 1–12, Vancouver, Canada. Association for Computational Linguistics.

Morales, M. R., Scherer, S., and Levitan, R. (2017b). OpenMM: An Open-Source Multimodal Feature Extraction Tool. In *Proceedings of Interspeech 2017*, pages 3354–3358, Stockholm, Sweden. ISCA.

Moreno, M. A., Jelenchick, L. A., Egan, K. G., Cox, E., Young, H., Gannon, K. E., and Becker, T. (2011). Feeling bad on facebook: depression disclosures by college students on a social networking site. *Depression and anxiety*, 28 6:447–55.

Mundt, J. C., Vogel, A. P., Feltner, D. E., and Lenderking, W. R. (2012). Vocal Acoustic Biomarkers of Depression Severity and Treatment Response. *Biological Psychiatry*, 72(7):580–587.

Nguyen, T., Phung, D., Dao, B., Venkatesh, S., and Berk, M. (2014). Affective and content analysis of online depression communities. *IEEE Transactions on Affective Computing*, 5(3):217–226.

Nuevo, R., Lehtinen, V., Reyna-Liberato, P. M., and Ayuso-Mateos, J. L. (2009). Usefulness of the beck depression inventory as a screening method for depression among the general population of finland. *Scandinavian journal of public health*, 37(1):28–34.

Østergaard, S. D., Jensen, S., and Bech, P. (2011). The heterogeneity of the depressive syndrome: when numbers get serious. *Acta Psychiatrica Scandinavica*, 124(6):495–496.

Owens, M., Herbert, J., Jones, P. B., Sahakian, B. J., Wilkinson, P. O., Dunn, V. J., Croudace, T. J., and Goodyer, I. M. (2014). Elevated morning cortisol is a stratified population-level biomarker for major depression in boys only with high depressive symptoms. *Proceedings of the National Academy of Sciences*, 111(9):3638–3643.

Oxman, T. E., Rosenberg, S. D., Schnurr, P. P., and Tucker, G. J. (1988). Diagnostic classification through content analysis of patients speech. *American Journal of Psychiatry*, 145(4):464–468.

Pakhomov, S., Chacon, D., Wicklund, M., and Gundel, J. (2011). Computerized assessment of syntactic complexity in alzheimers disease: a case study of iris murdochs writing. *Behavior research methods*, 43(1):136–144.

Parker, G., Hadzi-Pavlovic, D., Boyce, P. M., Wilhelm, K., Brodaty, H., Mitchell, P. B., Hickie, I. B., and Eyers, K. (1990). Classifying depression by mental state signs. *The British journal of psychiatry : the journal of mental science*, 157:55–65.

Pennebaker, J. W. (1997). Writing about emotional experiences as a therapeutic process. *Psychological science*, 8(3):162–166.

Pennebaker, J. W., Booth, R. J., and Francis, M. E. (2007). Linguistic inquiry and word count: Liwc [computer software]. *Austin, TX: liwc. net*.

Pérez Espinosa, H., Escalante, H. J., Villaseñor-Pineda, L., Montes-y Gómez, M., Pinto-Avedaño, D., and Reyez-Meza, V. (2014). Fusing affective dimensions and audio-visual features from segmented video for depression recognition: Inaoe-buap's participation at avec'14 challenge. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 49–55. ACM.

Pérez-Rosas, V., Abouelenien, M., Mihalcea, R., and Burzo, M. (2015). Deception detection using real-life trial data. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 59–66. ACM.

Pérez-Rosas, V., Mihalcea, R., and Morency, L.-P. (2013). Utterance-level multimodal sentiment analysis. In *ACL (1)*, pages 973–982.

Poria, S., Cambria, E., Bajpai, R., and Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society. IEEE Catalog No.: CFP11SRW-USB.

Pradhan, S., Elhadad, N., Chapman, W., Manandhar, S., and Savova, G. (2014). Semeval-2014 task 7: Analysis of clinical text. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, volume 199, pages 54–62.

Pyszczynski, T. and Greenberg, J. (1987). Self-regulatory perseveration and the depressive self-focusing style: a self-awareness theory of reactive depression. *Psychological bulletin*, 102(1):122.

Quatieri, T. F. and Malyska, N. (2012). Vocal-source biomarkers for depression: A link to psychomotor activity. In *Interspeech*.

Rafferty, A. N. and Manning, C. D. (2008). Parsing three german treebanks: Lexicalized and unlexicalized baselines. In *Proceedings of the Workshop on Parsing German*, pages 40–46. Association for Computational Linguistics.

Reece, A. G. and Danforth, C. M. (2016). Instagram photos reveal predictive markers of depression. *CoRR*, abs/1608.03282.

Resnik, P., Armstrong, W., Claudino, L., Nguyen, T., Nguyen, V.-A., and Boyd-Graber, J. (2015). Beyond lda: exploring supervised topic modeling for depression-related language in twitter. *NAACL HLT 2015*, page 99.

Ringeval, F., Schuller, B. W., Valstar, M. F., Gratch, J., Cowie, R., Scherer, S., Mozgai, S., Cummins, N., Schmi, M., and Pantic, M. (2017). Avec 2017 real-life depression, and aect recognition workshop and challenge.

Rosenberg, A. (2010). Autobi-a tool for automatic tobi annotation. In *INTERSPEECH*, pages 146–149.

Rosenberg, A. and Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. In *EMNLP-CoNLL*, volume 7, pages 410–420.

Rude, S., Gortner, E.-M., and Pennebaker, J. (2004). Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8):1121–1133.

Rush, A. J., Trivedi, M. H., Ibrahim, H. M., Carmody, T. J., Arnow, B., Klein, D. N., Markowitz, J. C., Ninan, P. T., Kornstein, S., Manber, R., et al. (2003). The 16-item quick inventory of depressive symptomatology (qids), clinician rating (qids-c), and self-report (qids-sr): a psychometric evaluation in patients with chronic major depression. *Biological psychiatry*, 54(5):573–583.

Saon, G. (2016). Word error rate: Recent advances in conversational speech recognition. `https://developer.ibm.com/watson/blog/2016/04/28/recent-advances-in-conversational-speech-recognition-2/`.

Schelde, J. T. (1998). Major depression: behavioral markers of depression and recovery. *The Journal of nervous and mental disease*, 186 3:133–40.

Scherer, K. R. (1986). Vocal affect expression: a review and a model for future research. *Psychological bulletin*, 99(2):143.

Scherer, S., Stratou, G., Gratch, J., and Morency, L.-P. (2013a). Investigating voice quality as a speaker-independent indicator of depression and ptsd. In *Interspeech*, pages 847–851.

Scherer, S., Stratou, G., Lucas, G., Mahmoud, M., Boberg, J., Gratch, J., Morency, L.-P., et al. (2014). Automatic audiovisual behavior descriptors for psychological disorder analysis. *Image and Vision Computing*, 32(10):648–658.

Scherer, S., Stratou, G., Mahmoud, M., Boberg, J., Gratch, J., Rizzo, A., and Morency, L.-P. (2013b). Automatic behavior descriptors for psychological disorder analysis. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–8. IEEE.

Scherer, S., Stratou, G., and Morency, L.-P. (2013c). Audiovisual behavior descriptors for depression assessment. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 135–140. ACM.

Schmidt, H. D., Shelton, R. C., and Duman, R. S. (2011). Functional biomarkers of depression: diagnosis, treatment, and pathophysiology. *Neuropsychopharmacology*, 36(12):2375–2394.

Schneider, D., Regenbogen, C., Kellermann, T., Finkelmeyer, A., Kohn, N., Derntl, B., Schneider, F., and Habel, U. (2012). Empathic behavioral and physiological responses to dynamic stimuli in depression. *Psychiatry research*, 200(2):294–305.

Schumann, I., Schneider, A., Kantert, C., Löwe, B., and Linde, K. (2012). Physicians attitudes, diagnostic process and barriers regarding depression diagnosis in primary care: a systematic review of qualitative studies. *Family practice*, 29(3):255–263.

Schwartz, H. A., Eichstaedt, J., Kern, M. L., Park, G., Sap, M., Stillwell, D., Kosinski, M., and Ungar, L. (2014). Towards assessing changes in degree of depression through facebook. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 118–125.

Segrin, C. (2000). Social skills deficits associated with depression. *Clinical psychology review*, 20 3:379–403.

Sharp, T. and Cowen, P. J. (2011). 5-ht and depression: is the glass half-full? *Current opinion in pharmacology*, 11(1):45–51.

Sidorov, M. and Minker, W. (2014). Emotion recognition and depression diagnosis by acoustic and visual features: A multimodal approach. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 81–86. ACM.

Skut, W., Krenn, B., Brants, T., and Uszkoreit, H. (1997). An annotation scheme for free word order languages. In *Proceedings of the Fifth Conference on Applied Natural Language Processing ANLP-97*, Washington, DC.

Sobin, C. and Sackeim, H. A. (1997). Psychomotor symptoms of depression. *The American journal of psychiatry*, 154 1:4–17.

Stassen, H. H., Kuny, S., and Hell, D. (1998). The speech analysis approach to determining onset of improvement under antidepressants. *European Neuropsychopharmacology*, 8(4):303–310.

Steiger, A. and Kimura, M. (2010). Wake and sleep eeg provide biomarkers in depression. *Journal of psychiatric research*, 44(4):242–252.

Stirman, S. W. and Pennebaker, J. W. (2001). Word use in the poetry of suicidal and nonsuicidal poets. *Psychosomatic Medicine*, 63(4):517–522.

Stodden, V. and Miguez, S. (2013). Best practices for computational science: Software infrastructure and environments for reproducible and extensible research. *SSRN*.

Sturim, D. E., Torres-Carrasquillo, P. A., Quatieri, T. F., Malyska, N., and McCree, A. (2011). Automatic detection of depression in speech using gaussian mixture modeling with factor analysis. In *Interspeech*, pages 2981–2984.

Sweeney, J. A., Strojwas, M., Mann, J. J., and Thase, M. E. . (1998). Prefrontal and cerebellar abnormalities in major depression: evidence from oculomotor studies. *Biological psychiatry*, 43 8:584–94.

Tausczik, Y. R. and Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Tomar, S. (2006). Converting video formats with ffmpeg. *Linux Journal*, 2006(146):10.

Trevino, A. C., Quatieri, T. F., and Malyska, N. (2011). Phonologically-based biomarkers for major depressive disorder. *EURASIP Journal on Advances in Signal Processing*, 2011(1):1–18.

Valderas, J. M., Starfield, B., Sibbald, B., Salisbury, C., and Roland, M. (2009). Defining comorbidity: implications for understanding health and health services. *The Annals of Family Medicine*, 7(4):357–363.

Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Cowie, R., and Pantic, M. (2016a). Summary for avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 1483–1484. ACM.

Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres Torres, M., Scherer, S., Stratou, G., Cowie, R., and Pantic, M. (2016b). Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 3–10. ACM.

Valstar, M., Schuller, B., Smith, K., Almaev, T., Eyben, F., Krajewski, J., Cowie, R., and Pantic, M. (2014). Avec 2014: 3d dimensional affect and depression recognition challenge. In *4th Audio/Visual Emotion Challenge Proc.*, pages 3–10. ACM.

Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., Schnieder, S., Cowie, R., and Pantic, M. (2013). Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In *3rd ACM international workshop on Audio/visual emotion challenge Proc.*, pages 3–10. ACM.

Venek, V., Scherer, S., Morency, L.-P., Rizzo, A., and Pestian, J. P. (2016). Adolescent suicidal risk assessment in clinician-patient interaction. *IEEE Transactions on Affective Computing.*

Villing, R., Timoney, J., and Ward, T. (2004). Automatic blind syllable segmentation for continuous speech.

Wagner, M. (2004). Prosody as a diagonalization of syntax. evidence from complex predicates. In *PROCEEDINGS-NELS*, volume 34, pages 587–602.

Wang, X., Zhang, C., Ji, Y., Sun, L., Wu, L., and Bao, Z. (2013). A depression detection model based on sentiment analysis in micro-blog social network. In *Trends and Applications in Knowledge Discovery and Data Mining*, pages 201–213. Springer.

Weide, R. L. (1998). The cmu pronouncing dictionary. *URL: http://www.speech.cs.cmu.edu/cgibin/cmudict.*

Williamson, J. R., Godoy, E., Cha, M., Schwarzentruber, A., Khorrami, P., Gwon, Y., Kung, H.-T., Dagli, C., and Quatieri, T. F. (2016). Detecting depression using vocal, facial and semantic communication cues. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 11–18. ACM.

Williamson, J. R., Quatieri, T. F., Helfer, B. S., Ciccarelli, G., and Mehta, D. D. (2014a). Vocal and facial biomarkers of depression based on motor incoordination and timing. In *4th Audio/Visual Emotion Challenge Proc.*, pages 65–72. ACM.

Williamson, J. R., Quatieri, T. F., Helfer, B. S., Ciccarelli, G., and Mehta, D. D. (2014b). Vocal and facial biomarkers of depression based on motor incoordination and timing. In *4th Audio/Visual Emotion Challenge Proc.*, AVEC '14, pages 65–72, New York, NY, USA. ACM.

Wolf, M., Horn, A. B., Mehl, M. R., Haug, S., Pennebaker, J. W., and Kordy, H. (2008). Computergestützte quantitative textanalyse: Äquivalenz und robustheit der deutschen version des linguistic inquiry and word count. *Diagnostica*, 54(2):85–98.

World Health Organization (2001). Mental health: a call for action by world health ministers. *Geneva: World Health Organization, Department of Mental Health and Substance Dependence*.

Yang, L., Jiang, D., He, L., Pei, E., Oveneke, M. C., and Sahli, H. (2016). Decision tree based depression classification from audio video and language information. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 89–96. ACM.

Yang, Y., Fairbairn, C., and Cohn, J. F. (2013). Detecting depression severity from vocal prosody. *Affective Computing, IEEE Transactions on*, 4(2):142–150.

Zinken, J., Zinken, K., Wilson, J. C., Butler, L., and Skinner, T. (2010). Analysis of syntax and word use to predict successful participation in guided self-help for anxiety and depression. *Psychiatry research*, 179(2):181–186.