Speech-based Automatic Depression Detection via Biomarkers Identification and Artificial Intelligence Approaches

Fuxiang Tao

Submitted in fulfilment of the requirements for the
Degree of Doctor of Philosophy

School of Engineering
College of Science and Engineering
University of Glasgow

August 2023

Abstract

Depression has become one of the most prevalent mental health issues, affecting more than 300 million people all over the world. However, due to factors such as limited medical resources and accessibility to health care, there are still a large number of patients undiagnosed. In addition, the traditional approaches to depression diagnosis have limitations because they are usually time-consuming, and depend on clinical experience that varies across different clinicians. From this perspective, the use of automatic depression detection can make the diagnosis process much faster and more accessible. In this thesis, we present the possibility of using speech for automatic depression detection. This is based on the findings in neuroscience that depressed patients have abnormal cognition mechanisms thus leading to the speech differs from that of healthy people. Therefore, in this thesis, we show two ways of benefiting from automatic depression detection, i.e., identifying speech markers of depression and constructing novel deep learning models to improve detection accuracy.

The identification of speech markers tries to capture measurable depression traces left in speech. From this perspective, speech markers such as *speech duration*, *pauses* and *correlation matrices* are proposed. Speech duration and pauses take speech fluency into account, while correlation matrices represent the relationship between acoustic features and aim at capturing psychomotor retardation in depressed patients. Experimental results demonstrate that these proposed markers are effective at improving the performance in recognizing depressed speakers. In addition, such markers show statistically significant differences between depressed patients and non-depressed individuals, which explains the possibility of using these markers for depression detection and further confirms that depression leaves detectable traces in speech.

In addition to the above, we propose an attention mechanism, Multi-local Attention (MLA), to emphasize depression-relevant information locally. Then we analyse the effectiveness of MLA on performance and efficiency. According to the experimental results, such a model can significantly improve performance and confidence in the detection while reducing the time required for recognition. Furthermore, we propose Cross-Data Multilevel Attention (CDMA) to emphasize different types of depression-relevant information, i.e., specific to each type of speech and common to both, by using multiple attention mechanisms. Experimental results demonstrate that the proposed model is effective to integrate different types of depression-relevant information in speech, improving the performance significantly for depression detection.

Acknowledgements

Throughout my four years in Glasgow pursuing a doctoral degree, there are countless people to be grateful for. Without their help, I wouldn't have been able to remain composed throughout this journey. I've encountered numerous challenges along the way, but these have only served to inspire me to face future difficulties with calm. This will be an unforgettable memory full of happiness, beauty, and value that will benefit me for the rest of my life.

Firstly, I want to thank my supervisor, Professor Alessandro Vinciarelli. He is an incredibly significant figure in my life. Without him, I wouldn't have even come to Glasgow. He didn't only teach me a vast amount of academic knowledge and how to view problems from various perspectives, providing me with academic inspiration, but also imparted many life lessons. From him, I learned how to be a kind, tolerant, polite, dedicated, responsible, and knowledgeable scholar. He has never been stingy in praising me to others, often promoting my academic achievements and recommending me to other scholars at academic functions. I am genuinely grateful to him. He has become the person I aspire to be.

Next, I would like to express my special thanks to Dr. Tanaya Guha, who frequently supports and encourages me. She is a wonderful person, often comforting me when my papers are rejected and validating my work, which is important to me. I want to thank Xuri Ge, and Professor Anna Esposito, who greatly assisted in publishing my papers. I would also like to thank Dr. Michele Sevegnani and Dr. Ke Yuan for acting as my annual progress reviewers, offering constructive feedback to help me complete my thesis more effectively. Thanks to Helen Border and Dr. Mireilla Bikanga Ada for giving me the opportunity to be a teaching assistant and for their support, from which I learned how to disseminate knowledge.

I want to express my gratitude to my family. Without their support and help, I wouldn't have been able to finish my studies. They always support and encourage me during difficult times, which enables me to persevere.

In addition to those mentioned above, I would also like to thank my friends who were or are in Glasgow. Thanks to my best roommate Tian Tian, and my very nice friends Wei Ma and Wen Sun, they always supported me when I was down. Thanks to Songpei Xu, Qiyuan Wang, Zejian Feng, Yingying Huang, and Weiyun Wang for sharing hotpot and playing SanGuoSha together. Thanks to Yaxiong Wu, Xiao Wang, and Yanni Ji for their support and encouragement, and to Moci Zheng and Zixiang Luo for the fun times we had together. Thank you, Qi Zhang. Thank

you to everyone who has helped me.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background on Depression

Depression, as a common mental health issue, can happen to anyone who has experienced excessively stressful events or trauma. In particular, people who have experienced abuse, significant losses, or other negative events in life are much likelier to be affected by depression. For example, extensive surveys of the population suggest that 16.2% of the adults in the USA experience at least one episode of Major Depression Disorder (MDD) during their life (R. C. Kessler et al., 2003). According to a large-scale analysis (Vos et al., 2016), the number of individuals with depression increased by 18.4% between 2005 and 2015. In 2015, the World Health Organization states that depression affected 4.4% of the world's population, corresponding to 322 million people (WHO, 2017). In particular, according to the statistics from the Center for Disease Control, adolescents ages 13–18 have experienced one of the depression symptoms (Kann et al., 2018). Furthermore, a population screening suggests that depression incidence increased by 37% from 2005 to 2014 (Mojtabai et al., 2016). Therefore, on current trends, depression is likely to become the most common mental illness by 2030. (Mathers & Loncar, 2006). The COVID-19 pandemics further aggravated such a serious situation, resulting in an even greater number of patients (Bueno-Notivol et al., 2021). There was an estimation of an additional 53.2 million cases of major depressive disorder per 100,000 population globally due to COVID-19, corresponding to an increase of 27.6% (Santomauro et al., 2021).

At the individual level, depression increases the risks of suicide and suicide-caused disability. The World Health Organization estimates that 800,000 suicides and 16,000,000 suicide attempts happened globally in 2016 (WHO, 2017). More than half of them appear to be associated with depression (R. C. Kessler et al., 1994), and approximately 66% of individuals diagnosed with depression exhibit an increased likelihood of engaging in suicide attempts during their lives (McLaughlin, 2011). Depression is one of the main causes of disability among people above 5 years of age (Gotlib & Hammen, 2008). According to the World Health Organization, depression led to a total of over 50 million years lived with disability over the world,

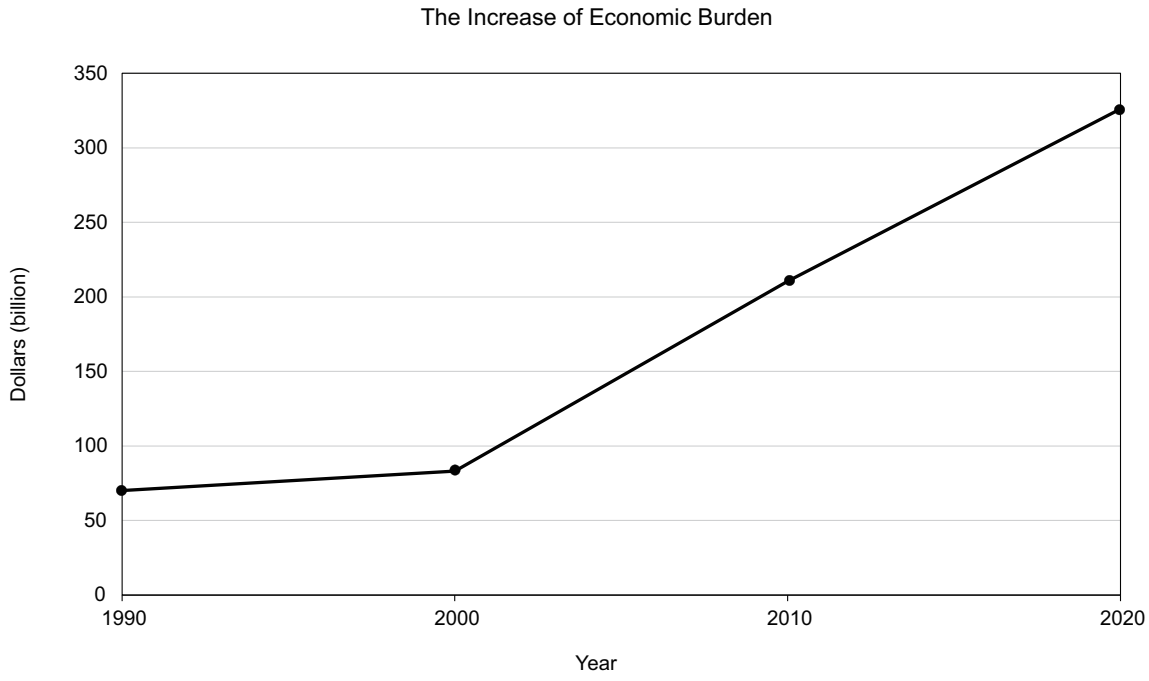

Figure 1.1: This figure shows the annual costs including medical expenses, productivity loss, etc., associated with depression in the USA from 1990 to 2020.

corresponding to 7.5% of all years lived with disability (WHO, 2017). At present, depressive disorders have been ranked as the top contributors to non-fatal health losses, the single largest factor. Consequently, depression places substantial burdens on individuals, families, and society as a whole.

As a consequence of the above, depression becomes a major economic and societal burden at the family or society level. For example, in Canada, the average medical costs for depressed people are 3.5 times higher than those for non-depressed ones (Tanner et al., 2020). In addition, the economic burden of adults with depression has increased significantly over time. For instance, in the USA, the annual costs associated with depression (medical expenses, productivity loss, etc.) increased as follows: 70 billion dollars per year in 1990 (P. E. Greenberg et al., 1993); 83.1 billion dollars in 2000 (P. E. Greenberg et al., 2003); 173.2 billion dollars in 2005; and 210.5 billion dollars in 2010 (P. E. Greenberg et al., 2015). In 2020, this number went up to 326.2 billion dollars per year (P. E. Greenberg et al., 2021). This corresponds to a 366% increase even (see Figure 1.1). The economic burden on society is heavier in developing countries. This is mainly because over 75% of individuals living in developing countries, get too low income to access treatment, despite the availability of established and effective treatments for mental disorders (Evans-Lacko et al., 2018).

Similar to the estimate of depression in prevalence developing countries, the figures above are still likely to be an underestimate because, globally, only half of the patients obtain medical attention and, furthermore, only 21% of them undergo adequate treatment (R. C. Kessler et al.,

2003). In other words, while being a serious pathology with highly negative consequences, depression tends to remain undetected or to be poorly treated. One possible reason for such a situation is that a large number of cases go undiagnosed due to limited healthcare access in many countries (Hasin et al., 2018; Williams et al., 2017).

1.2 Depression Diagnosis

In current clinical research, depression assessment is typically determined through various questionnaires, including the Hamilton Rating Scale for Depression (HAM-D) (Hamilton, 1986), Patient Health Questionnaire (PHQ) (Kroenke et al., 2001), Beck Depression Inventory-II (BDI-II) (Beck et al., 1996), Diagnostic and Statistical Manual of Mental Disorders-IV (DSM-IV) (APA, 2013), Quick Inventory of Depressive Symptomatology (Rush et al., 2003), Montgomery Åsberg Depression Rating Scale (MADRS) (Montgomery & Åsberg, 1979) and Youth Mania Rating Scale (YMRS) (Young et al., 1978). However, these traditional assessments heavily rely on subjective judgement based on the experience and skills of clinicians. According to the epidemiological observations, these traditional assessments lead to missed diagnosis in which only 26.4% depressed patients undergo diagnosis (Faisal-Cury et al., 2022) and 21% patients undergo adequate treatment (R. C. Kessler, 2003). In other words, while being a serious pathology with highly nediagnosisgative consequences, depression tends to remain undetected or to be poorly treated via traditional approaches. One possible reason for such a situation is that the first line of intervention against depression is not led by psychiatrists, specialised and experienced in the diagnosis of mental health issues, but by General Practitioners (GPs), doctors expected to deal with common pathologies, but to refer to specialists in case of more serious problems. In particular, due to the difficulties in diagnosing depression, the accuracy of GPs has been shown to range between 57.9% and 73.1%, thus leaving a large number of cases undetected (Mitchell et al., 2009).

Another possible reason for the situation is that there is an inconsistency between mood and cases of clinical trials (H. Yang et al., 2005). For example, the effects of self-report bias are ubiquitous in clinical depression diagnosis, with males tending to minimize their depressive symptoms more than females (Hunt et al., 2003), therefore, these lead to difficulties in the interpretation of the results related to depression. On the one hand, the methodological pitfalls and experimental design factors, including patient selection and enrollment pressures, can also contribute to inaccurate and unreliable clinical findings (Demitrack et al., 1998). On the other hand, there are potential rater biases in depression diagnosis. For instance, effect sizes based on clinician ratings were significantly larger than patient-based ratings (R. P. Greenberg et al., 1992).

Recently, more and more researchers started to investigate the detection of depression automatically by analyzing data obtained from multiple sources. Neuroimaging encompasses a

wide range of non-invasive approaches that offer insights into neural mechanisms related to depression from the perspective of brain structure and function. Neuroimaging technique, such as Functional Magnetic Resonance Imaging (fMRI), electroencephalogram (EEG), magnetoencephalography (MEG), have been applied and played an important role in the investigation of depression. For example, researchers observed a decreased plasma brain-derived neurotrophic factor in untreated major depressive disorder patients (Lee et al., 2007), and suicide led by depression is related to lower Cerebrospinal fluid levels of 5-hydroxyindolacetic acid in the depressed brain (Placidi et al., 2001), and therefore, these can be considered as biomarkers in depression diagnosis. In addition to the potential biomarkers found in the brain, researchers have also found similar traces in blood and nervous systems. On the one hand, abnormal levels of cytokines and 5-HTT have been revealed as an immune response to the central nervous system in the pathogenesis of depression (Tsao et al., 2006). On the other hand, researchers found that depressed patients have abnormal changes in hypothalamus-pituitary-adrenocortical system regulation assessed with repeated dexamethasone/corticotropin releasing hormone tests (Ising et al., 2007). These potential biomarkers may predict clinical outcomes in depression diagnosis. One of the main issues in addressing the problems above is that depression diagnosis is a difficult, expensive and time-consuming process. This requires clinicians to take a lot of time and effort to make correct decisions towards individuals with depression. In addition, the diagnoses mentioned above are invasive, requiring laboratory settings, and therefore making the diagnosis process more complicated.

To this end, researchers are developing automatic approaches capable of identifying, at least to a certain extent, people affected by depression. However, the accuracy of the current automatic depression detection approach is still limited and difficult to interpret. The focus of the thesis is on possible explainable depression markers in speech, i.e., on automatically detectable traces of the pathology that can contribute to improving the performance of depression detection approaches. The main motivation for investigating speech is that we believe the process of speech production differs between depressed and healthy speakers as we will demonstrate in Chapter 2, and speech is easier to analyze in real-world situations. The main motivation for focusing on markers is that these cannot be easily controlled by patients that, in some cases, try to conceal the pathology to escape the stigma associated with it. In this respect, markers can complement questionnaires or other observational clinical approaches, often affected by subjective biases, with more objective measurements. Another aspect of this thesis is using machine learning and deep learning approaches to emphasize depression-related information in supporting depression diagnosis.

1.3 Thesis Statements

The key-idea underlying this thesis is that depression leaves measurable traces in speech which can help to distinguish between depressed and non-depressed speakers. The first step in discovering these traces is to establish a more reliable depression speech corpus. Specifically, this thesis posits that multiple depression-related traces exist, such as those that could be found in temporal and acoustic properties. The potential implications of analyzing these traces include the identification of speech markers that may contribute to clinical depression diagnosis. To analyze these traces and extract depression-relevant information, we also propose the use of machine learning or deep learning approaches. Such methodologies can uncover potential patterns or relationships related to depression that may not be easily discernible through traditional statistical methods, thereby enhancing the performance in distinguishing between depressed and non-depressed speakers. In particular, we postulate that utilizing attention mechanisms based on deep learning allows emphasising depression-related information, and further improving recognition performance and efficiency in depression detection. Lastly, we propose that utilizing multiple attention mechanisms in different types of speech (read and spontaneous) can capture depression-related information, both specific and common in different types of speech. The combination of this information can further enhance performance in distinguishing between depressed and non-depressed speakers.

In summary, by analysing these measurable traces in speech, our goal is to provide a more objective approach to depression diagnosis. This approach has the potential to complement current clinical practices and streamline the diagnosis process in a much simpler way (e.g., through the use of speech). As a result, it can reduce the reliance on subjective evaluations, and enhance the overall accuracy and efficiency of depression detection. In the long run, our findings in this thesis may contribute to the development of more effective diagnostic tools for those living with depression.

1.4 Thesis Structures and Contributions

The remainder of the thesis is organized as follows:

- Chapter 2 provides background about depression, including cognition changes in the depressed brain, speech production changes related to depression and related work for automatic depression detection. First, we introduce the abnormal cognition mechanisms in the depressed brain, and how such mechanisms are linked to speech, from the perspective of language and cognition. Second, we describe the changes in speech production among depressed patients, resulting from the changes in cognition and motor muscular actions. In the last part of this chapter, we review the techniques in automatic depression detection including the type of data and deep learning approaches for detection.

- Chapter 3 describes a novel benchmark, the Androids Corpus, used in this thesis for speech-based depression detection which consists of 118 recordings of depressed and control participants. We first review the publicly available speech datasets for depression detection, and make a comparison with ours (the Androids Corpus). Then we describe the details of the corpus including tasks description (i.e., Read Task and Interview Task), data distribution and demographic information about participants (age, gender, education level). In the last part of the chapter, we provide the baseline approaches including feature extraction, depression recognition, and results for both read and spontaneous speech.

  Contribution: providing a detailed description of the new dataset for depression detection. This publicly available corpus provides preliminary experimental studies with the proposed experimental protocol including types of speech material, data collection protocol, speaker-independent protocol, clear experimental setup and results of applied baseline systems. Compared to the other available datasets, this corpus includes more participants with read and spontaneous speech from the same speaker. Therefore, this allows to investigate multiple research questions in the following chapters, such as expanding proposed approaches on different types of speech to make a comparison or interaction.

- Chapter 4 presents two timing-based speech features obtained by using computational linguistic and social signal processing approaches, namely *duration of speech* and *silence*. Such features show significant differences between depressed and non-depressed speakers, and can be used as effective markers for depression detection. We first review the related work for the identification of timing-based markers. Then we extract these speech features and examine their effectiveness for depression detection in both read and spontaneous speech. Later in the chapter, the experiments show that such markers benefit depression detection, and are robust in both types of speech.

  Contribution: proposing an approach that creates a voice-based depression classifier with proposed speech features. In particular, defining and applying a controlled method to obtain a lean classifier is unlikely to be overfitting. From a pragmatic point of view, the proposed approach provides a way to evaluate the effectiveness of any proposed speech features by combining computational paralinguistics and social signal processing approaches.

- Chapter 5 presents another marker to characterize the relationship between acoustical properties, namely *correlation matrices*. We first review previous related work on using speech feature vectors and correlation structures. Then we propose the approach for correlation matrices representation and apply these matrices in depression recognition. We next propose *stability* to measure the changes in the correlation matrices over time. In the last part of the chapter, we show the experimental results which prove the effectiveness of using correlation matrices in both read and spontaneous speech for depression detection.

  Contribution: the first contribution is that it presents a novel approach by comparing

SVM and LSTM models when fed with feature vectors versus correlation matrices. The second contribution is that it demonstrates that changes in correlation patterns over time can serve as a depression marker, which is a novel finding. The last contribution is that we introduce a measurement (stability) as an explanation to the points mentioned above, that is correlation matrices convey more variance that can help to distinguish depressed and non-depressed speakers, which is of benefit in depression detection.

- Chapter 6 introduces a novel model with an attention mechanism, the *Multi-Local Attention*, that aims at emphasizing depression-related information for depression detection. We first describe the architecture of the proposed model, subsequently, we propose a *confidence score* to measure how confident the model is about its outcomes. Next, we show the performance of depression detection with the proposed approach and make comparisons to baselines and state-of-the-art approaches. Finally, we examine Multi-Local Attention in other two respects, i.e., the confidence and the time used for prediction.

  Contribution: proposing a novel attention mechanism that enhances the accuracy of the LSTM-based detector. The proposed approach outperforms state-of-the-art not only in performance but also in other two aspects: 1) it significantly improves the confidence in the outcomes classified correctly, meaning it allows the model to make recognition for the cases evident enough, in the meanwhile, it improves the diagnosis efficiency of doctors and allows them to focus on more ambiguous cases; 2) it significantly reduces the time required for testing a potential patient.

- Chapter 7 proposes a novel model, Cross-Data Multilevel Attention (CDMA), with multiple attention mechanisms to explicitly emphasize all the depression-relevant information available in the data, whether it is specific to each type of speech or common to both of them. We first review the previous work using read and spontaneous speech and attention mechanisms for depression detection. Then we introduce the architecture of the proposed model and the interactions of multiple attention mechanisms for depression detection. Finally, we analyse the performance for each step and make comparisons by adding one component at each time.

  Contribution: providing another novel architecture that takes read and spontaneous speech into account jointly. This approach is designed for depression detection, but it can be used for other data consisting of pairs of items that share a common label in a classification or regression problem. Another contribution of this chapter is the experimental evidence that read and spontaneous speech carry both specific and common information. By combining such information, it is sufficiently diverse to enhance the performance of a detection approach, outperforming the state-of-the-art approaches as well as baselines.

1.5 Supporting Publications

Most of the materials presented in this thesis are built on the publications considered by various international conferences and journals, as follows:

- **Fuxiang Tao**, Anna Esposito, and Alessandro Vinciarelli. "The Androids Corpus: A New Publicly Available Benchmark for Speech Based Depression Detection" (**INTERSPEECH 2023, published**) (**Chapter 3**)
- **Fuxiang Tao**, Anna Esposito, and Alessandro Vinciarelli. "Spotting the Traces of Depression in Read Speech: An Approach Based on Computational Paralinguistics and Social Signal Processing." (**INTERSPEECH 2020, published**) (**Chapter 4**)
- **Fuxiang Tao**, Wei Ma, Xuri Ge, Anna Esposito, and Alessandro Vinciarelli. "The Relationship Between Speech Features Changes When You Get Depressed: Feature Correlations for Improving Speed and Performance of Depression Detection" The 30th International Conference on Neural Information Processing (**under review**) (**Chapter 5**)
- **Fuxiang Tao**, Xuri Ge, Wei Ma, Anna Esposito, and Alessandro Vinciarelli. "Multi-Local Attention for Speech-Based Depression Detection" IEEE International Conference on Acoustics, Speech and Signal Processing (**ICASSP 2023, published**) (**Chapter 6**)
- **Fuxiang Tao**, Xuri Ge, Wei Ma, Anna Esposito, and Alessandro Vinciarelli. "Cross-Data Multilevel Attention Mechanisms for Depression Detection: Analyzing the Interplay Between Read and Spontaneous Speech" (**to be submitted a journal**) (**Chapter 7**)

Chapter 2

Background

In Chapter 1, we introduced the societal effects of depression including health and economic burdens, and the limitations of depression diagnosis in clinical. To alleviate those limitations such as high expenses, time costs, complex diagnostic processes, etc. We have proposed speech-based automatic depression detection from two aspects, one is to identify effective speech markers to help the process of detection, the other is to use deep learning approaches with attention mechanisms to detect depressed speakers. In this chapter, we first provide the reasons for proposing the use of speech for depression detection from two points of view. One is the abnormal cognition mechanisms in the depressed brain and how these mechanisms affect speech processing. The other one is how these changes in the brain affect speech production. These two aspects build the theoretical basis of this thesis. In the last part of this chapter, we provide an overview of automatic depression detection.

2.1 Cognition in Depressed Speakers

Depression, as a mental disorder, is often characterized by a variety of distinctive behaviors. These include persistent feelings of sadness, a loss of interest in enjoyable activities, a noticeable decrease in energy, and difficulty with sleep (APA, 2013). Such manifestations are not merely symptoms and have been associated with tangible changes in cognition among depressed people. Understanding the influence of depression on mental and emotional states largely relies on the depression cognitive model, first proposed by Beck about half a century ago (Dobson, 1989). This model set the groundwork for a better understanding of the factors involved in depressive episodes and paved the way for the integration of neurobiological findings into the understanding of depression. With the help of advanced neuroimaging and biological techniques that have been developed over the years, we are now able to identify and understand such changes in cognition resulting from the changes in brain structures and functions (Ritchey et al., 2011). There is an increasing body of evidence suggesting that the behaviors associated with depression are linked to these alterations (Disner et al., 2011). This will affect the speech of depressed patients in two

ways, one is to affect speech through abnormal language production mechanisms, and the other is to affect speech indirectly through emotion (see below).

2.1.1 Depression Affects Language-related Areas of Brain

The frontal lobe is one of the four primary lobes in the brain, located at its forefront. This area executes sophisticated cognitive functions, for example, language. It has been observed that this region undergoes functional alterations in individuals living with depression (Disner et al., 2011). Furthermore, this lobe plays a pivotal role in various aspects of language processing (Alexander et al., 1989). Any observed dysfunctions in this area might explain the observed changes in the language use among those afflicted with depression.

One of the critical functions of the frontal lobe is muscle control and movement. As a result, the functionality of the glottis, a muscle critical to speech production, can be impacted among depressed individuals due to alterations of functioning in the frontal lobe. This could potentially lead to abnormal speech production in such individuals, a subject which we delve into in greater detail in Section 2.2.

Despite its relatively large size, accounting for approximately one-third of the surface area of each hemisphere, the functions of the frontal lobe are remarkably complex. This complexity has sparked interest among researchers, prompting them to propose sophisticated experimental techniques to explore how this region influences language processing. The study (Binder et al., 2003), for instance, employed event-related functional magnetic resonance imaging (fMRI) to demonstrate stronger word activation occurring in a distributed, left hemisphere network previously associated with semantic processing. This network includes the prefrontal cortex (part of the frontal lobe) and the angular gyrus (found in the parietal lobe), both of which play crucial roles in language processing (see Sections below for more details).

Broca's Area and Wernicke's area

Another region within the frontal lobe of particular interest is the left inferior frontal gyrus, also known as Broca's area. This area was first associated with language processing by Pierre Paul Broca, who observed speech impairments in two patients following injury to this brain region (Kennison, 2013; Dronkers et al., 2007). Subsequent extensive studies involving more patients with similar injuries corroborated the findings of Broca, suggesting that these patients exhibited significantly slower responses to target words. This further confirms the role of this area in word comprehension (Bedny et al., 2007).

Moreover, studies involving healthy participants have also attested to the role of the left inferior frontal gyrus in language processing. For instance, one study demonstrated that activation during a sentence-matching task primarily arose from the left inferior frontal gyrus (D'Arcy et al., 2004). A similar activation pattern was observed during phonologic (rhyme detection) and

semantic categorization tasks in fMRI experiments (Seghier et al., 2004).

Researchers found changes in this region among depressed people, for example, according to preliminary studies in neuroscience, depression is associated with a general dysfunction in the left frontal lobe (Davidson et al., 2002; W. Heller & Nitscke, 1997). Specifically, a recent fMRI-based study focusing on bipolar disorder (a type of depression) revealed functional dysconnectivity in the left inferior frontal gyrus of patients (Roberts et al., 2017).

In addition to Broca's area, another important area related to language is Wernicke's area, also called Wernicke's speech area. This area is particularly crucial for the comprehension of both written and spoken language, providing a complementary function to Broca's area, which primarily orchestrates language production. A remarkable discovery has been made concerning the delayed activation pattern of the Wernicke's area in individuals living with depression during semantic encoding of words (Abdullaev et al., 2002). It was found that these individuals required the activation of additional brain regions, including the right lateral prefrontal cortex, to perform semantic processing. This suggests that depression could potentially influence both the production and comprehension of language.

Prefrontal Cortex

The dorsolateral prefrontal cortex (DLPFC), a crucial region of the frontal lobe, has long been linked to higher-order cognitive functions. A large number of studies implied that depressed individuals exhibit distinct variations in their DLPFC compared to non-depressed controls. For instance, a voxel-based morphometry neuroimaging study has indicated that depressed individuals exhibit reduced gray matter volume in the DLPFC, suggesting a structural alteration in the brain (Li et al., 2010). On a functional level, another neuroimaging review study proposed that the DLPFC exhibits reduced activity in depressed brains (Gotlib & Hamilton, 2008). Notably, this study suggested that such underactivation might not occur in isolation but might be part of a broader pattern of functional connectivity characterizing brain functioning in depression.

Researchers have started to explore functional connectivity in depression from a neural systems perspective. Functional connectivity is defined as the temporal correlation of neurophysiological events that are spatially distant (Friston et al., 1994). In other words, two regions are considered functionally connected if their activity measures show a statistical correlation. Based on this, researchers have proposed the use of effective connectivity, which accounts for both structural and functional connectivity (Friston, 2011). The concept of effective connectivity in neuroscience extends beyond the basic notion of functional connectivity, it incorporates the directional influence that one neural element exerts over another, providing a more comprehensive and dynamic understanding of brain interactions. A recent neuroimaging study investigating such effective connectivity networks in the brain has revealed an abnormal pattern in depressed patients (Rolls et al., 2018). Crucially, it was observed that individuals with depression showed increased activity in the lateral orbitofrontal cortex, coupled with decreased effective connectiv-

ity to and from cortical language-related areas (Rolls et al., 2018). This finding corroborates the idea that alterations in prefrontal cortex function can impact language processing in the brain from a neurological perspective.

In addition, many studies have demonstrated a relationship between the DLPFC and language, particularly verbal fluency. For instance, a study that explored brain activation during verbal fluency tasks suggested that these tasks engage frontal and frontotemporal brain regions (Elfgren & Risberg, 1998). This pattern of activation was also identified in healthy children and adults during verbal fluency tasks, as determined by another fMRI study (Gaillard et al., 2000). These observations led several researchers to postulate that verbal fluency might be inherently tied to the functions of the frontal lobe. As a result, they suggested that tests of verbal fluency could potentially serve as indicators or assessments of frontal lobe functionality (Ravnkilde et al., 2002). This proposition is further supported by other studies evaluating phonemic verbal fluency in subjects living with depression, which generally revealed a bilateral hypoactivation within the frontal lobe (Klumpp & Deldin, 2010). From a behavioral perspective, evidence also points towards a noteworthy decline in task performance among patients with major depression relative to healthy controls during an overt and continuous semantic verbal fluency task (Backes et al., 2014). Taking into account the finding that alterations in the functionality of these specific areas are observed in patients with depression, it is plausible to infer that such changes could potentially impact the speech patterns of these individuals.

Angular Gyrus

A comprehensive review consisting of 275 neuroimaging investigations suggests a critical involvement of the left angular gyrus in processing language (Cabeza & Nyberg, 2000). This implication is further substantiated by an extensive fMRI analysis involving 901 participants, which revealed that depression influences the functional connectivity of several brain regions, particularly the left angular gyrus (Cheng et al., 2016). This region exhibited altered functional connectivity in the context of depression, which could potentially explain the observed inferior performance in semantic verbal fluency tasks among individuals living with depression (Cheng et al., 2016).

From the perspective of functional connectivity, the prefrontal cortex and angular gyrus are critical hubs in the default mode network (a large-scale network) in the brain. These large-scale brain networks are essentially collections of distributed brain regions exhibiting functional interplay. According to the points from neuroscientists, cognitive tasks are enacted not by isolated brain regions, but by dynamic networks comprising several discrete regions. Importantly, these key-regions are instrumental in executing cognitive tasks, such as reading a story (Regev et al., 2013).

However, pathologies like depression have been linked with abnormal overactivity within these critical DMN regions (Sheline et al., 2010). A study using dynamic functional connectivity

analysis to assess the temporal stability of these connections within the DMN found reduced stability over time (Wise et al., 2017). This finding was further corroborated in an independent sample. From the perspective of functional connectivity networks, patients with depression are more likely to be affected by the pathology to produce abnormal speech.

Clinical observations further strengthen these findings. Multiple studies have reported improvements in the speech of depressed patients following treatment. For example, a behavioral study involving 105 patients with major depressive disorder (MDD) observed significant changes in speech production patterns in response to treatment (Mundt et al., 2012). Similarly, a large meta-analysis revealed that MDD patients performed poorly in both phonemic and semantic verbal fluency tasks compared to healthy controls (Wagner et al., 2012). However, such worse performances of depressed patients were observed to improve over the course of treatment, which confirmed the negative effects of depression on speech production.

These findings show that depression impacts speech production from the perspective of neuroscience, potentially leading to observing differences in speech patterns between depressed individuals and healthy controls. This may contribute to developing techniques for automatic depression detection in the future.

2.1.2 Cognition Mechanisms in Depressed Brain

In addition to the impact on the regions related to language processing, depression also affects other cognition mechanisms that indirectly affect speech processing in the brain. In this subsection, we introduce two main cognition mechanisms interacting with depression, referred to as *biased attention* and *biased information processing*.

Biased Attention

A most common behavioral manifestation of depression is rumination, involving a persistent focus on negative content in the past or at present, resulting in emotional distress (Nolen-Hoeksema et al., 2008). This view is confirmed through a cognitive perspective, with substantial evidence suggesting that individuals with depression exhibit a particularly selective attention and preferential recall towards negative information (Mathews & MacLeod, 2005). Generally, this attention bias is not unique to depressed individuals, for example, non-depressed individuals tend to exhibit biased attention toward positive stimuli, such as uplifting or pleasurable content (Gotlib et al., 2004). However, a clinical study, using eye-tracking methodologies, has extended these findings to show that depressed individuals tend to spend more time on negative emotional stimuli, often neglecting the positive ones compared to healthy people (Kellough et al., 2008).

One plausible explanation for this cognitive behavior is the alterations in the frontal lobe. In addition to its role in language processing (see above), the frontal lobe also plays an impor-

tant role in human attention. Specifically, an fMRI-based neuroimaging study suggested that the disengagement of attention requires top-down intervention from higher-order brain cortical areas, including the prefrontal cortex (Corbetta et al., 1998). Given the alterations in the frontal lobe observed in depressed individuals (see above), it is expected that these individuals would exhibit different patterns of attention when compared to healthy people. According to Beck's depression cognitive model, biased attention in depressed individuals are due to the difficulty in disengaging from negative stimuli (Dobson, 1989). This difficulty is attributed to impairment in certain brain regions specific to depression, which lead to deficits in inhibiting attention towards negative stimuli (Disner et al., 2011).

Researchers have started to use neuroimaging techniques to understand such biased attention mechanisms in depression from the perspective of cognition. In particular, the ventrolateral prefrontal cortex and DLPFC are two crucial parts of the prefrontal cortex that interest researchers. In general, the ventrolateral prefrontal cortex is considered to be associated with control over stimulus selection, while the DLPFC is considered to be related to executive functioning (Passarotti et al., 2009). However, depressed individuals exhibit a failure to get feedback from the DLPFC, indicating impaired top-down cognitive control (Fales et al., 2008). In addition, another neuroimaging study further revealed the relationship between depression and cognitive control of emotion cues, in which individuals with depression have difficulty engaging neural regions that facilitate cognitive control, especially for emotional information, including the ventrolateral prefrontal cortex (Beevers et al., 2010). Therefore, these abnormalities in top-down cognitive control mechanisms contribute to the difficulty in disengaging from negative stimuli in depressed patients.

The anterior cingulate cortex (ACC), a brain region in the limbic system, is also considered to be associated with cognitive control. According to the finding in neuroscience, ACC is a key node in the circuit involved in regulating both cognitive and emotional processing (Bush et al., 2000). A neuroimaging study suggested that inhibition within negative stimuli recruits a distinct set of brain regions that includes ACC (Shafritz et al., 2006). As we mentioned in the above subsections, dynamic functional connectivity in DMN in the depressed brain tends to be unstable (see above). Importantly, ACC is also a crucial region in the DMN, meaning that it is expected to exhibit a unique pattern of activity in depressed individuals compared to healthy adults. According to a neuroimaging study related to depression, only negative stimuli can activate ACC in depressed individuals, although both positive and negative stimuli can activate ACC in non-depressed individuals (Eugène et al., 2010; Mitterschiffthaler et al., 2008). This differential response provides a complementary explanation for why depressed patients struggle to disengage from negative emotions.

In conclusion, alterations in the ACC and prefrontal cortex contribute to abnormal top-down regulation, leading to biased attention towards negative stimuli in depressed individuals. As a result, this biased attention to the input into the brain may potentially lead to biased information

processing for depressed patients.

Biased Information Processing

Beck's cognitive model of depression is a landmark that provides an understanding of the biased cognitive mechanisms that underpin depression (Beck, 2008). One of its most important features is the concept of biased information acquisition and processing, suggesting that both internal and external environmental stimuli activate and influence the information processing pathways in depressed patients. In line with previous sections, this suggests that depressed patients tend to focus on negative elements, resulting in the brain being trapped in the loop of negativity that is continually processing negative information.

From a neuroscience perspective, individuals with depression exhibit distinct patterns of biased information processing when encoding, organizing, and retrieving information (Disner et al., 2011). For example, depressed patients show decreased sensitivity to their positive emotions (Watson et al., 1988), in other words, a loss of pleasure. The main explanation for such a situation is the decrease of dopamine secretion, because dopamine is linked with the experience of pleasure. In particular, several neuroimaging studies suggested that a top-down control from the prefrontal cortex triggers dopamine release, increasing brain activity in response to rewarding stimuli and successfully regulating emotion by activity (Del Arco & Mora, 2008; Wager et al., 2008). However, this process can be impacted in individuals with depression due to the alterations of the prefrontal cortex (see subsections above).

Another critical function of the prefrontal cortex is emotion regulation, and a correlation has been demonstrated between specific patterns of the activity in the prefrontal cortex and emotion regulation (S. H. Kim & Hamann, 2007). The prefrontal cortex has been revealed with diminished maintenance of brain activation, which reflects a decreased capacity to keep positive emotion (A. S. Heller et al., 2009). This can be a key feature of depression that limits response to positive stimuli, which leads to biased information processing.

Depression is a mental disorder that can lead to changes in emotions, thus the brain regions related to emotion processing are the focus of researchers. The perception of emotional stimuli depends on the amygdala, two almond-shaped clusters of nuclei located in the temporal lobes of the brain. This brain region provides emotional responses and contributes to the interpretation of the emotional quality of stimuli. However, many studies suggested that there are structural and functional abnormalities in the amygdala in the depressed brain, such as increased neurophysiological activity, abnormal metabolism, and reduced grey matter volume (Drevets, 2001). In particular, depressed patients exhibit sustained amygdala activities in response to processing negative information, and such activities are more intense and longer than those in healthy controls (Siegle et al., 2002). Importantly, such excessive activity in the amygdala still persists in depressed individuals even after the removal of negative stimuli (Schaefer et al., 2002). These neuroscience findings reveal the impact of amygdala alterations on the abnormal processing of

negative emotional information in depression.

A comprehensive meta-analysis involving 385 studies found a negative correlation between amygdala activity and left DLPFC function when processing emotional information in healthy individuals (Costafreda et al., 2008). In general, the DLPFC exerts an indirect inhibitory regulation on the amygdala (J. LeDoux, 2007; J. E. LeDoux, 2000). However, the inhibitory deficits in the frontal areas can affect amygdala activity. For example, the decreased top-down regulation, from bilateral DLPFC to the amygdala, generates abnormal reactivity of the amygdala (Fales et al., 2008). Moreover, the hyperactivity of the amygdala in depressed individuals, acting as a bottom-up signal, influences higher cortical structures, contributing to the bias towards negative emotional processing, even at a subconscious level (Victor et al., 2010). This implies a tendency of the depressed brain to amplify negative emotional responses.

Therefore, alterations in the function and structure of the prefrontal cortex in depressed individuals, contribute to decreased cognitive control and amygdala hyperactivity. This is due to the dysfunction of both top-down and bottom-up regulation, leading to dysfunctions in emotional processing in the depressed brain. Such dysfunctions have an impact on cognitive processes, particularly in regard to attention and information processing, which is significant and has been shown to affect alterations in linguistic content (see next section for more details).

In conclusion, it is apparent that alterations in specific brain regions associated with language processing can directly influence speech production. In the meanwhile, changes in other brain regions can generate attention and information processing biases in individuals with depression. The attention bias can skew the focus of an individual towards negative content, thus further aggravating the biased information processing within the depressed brain. As a result, this predisposition towards negativity indirectly affects the speech content of depressed individuals, causing a tendency towards more negative words (see below), as shown in Figure 2.1.

2.2 Different Speech Production in Depression

Speech production, a common but complex process, is initiated with cognitive planning, in which the message going to be spoken is formulated and sets up phonetic and prosodic information in the brain of the speaker. Then such information is conveyed by motoric muscular actions (e.g. the control of the vocal tract), and this information is regulated by articulators for voice generation (Singh, 2019; Sataloff et al., 2007). Prosodic information depends on the style and rhythm of speech.

In the previous sections of this chapter, we reviewed the various neurobiological alterations that occur within the depressed brain, and introduced the multiple cognitive impairments especially for the mechanisms of language processing in depression (see Section 2.1). These alternations are expected to affect the speech production of depressed patients. In the current section, we will introduce speech from the perspectives of motor actions and voice. We will

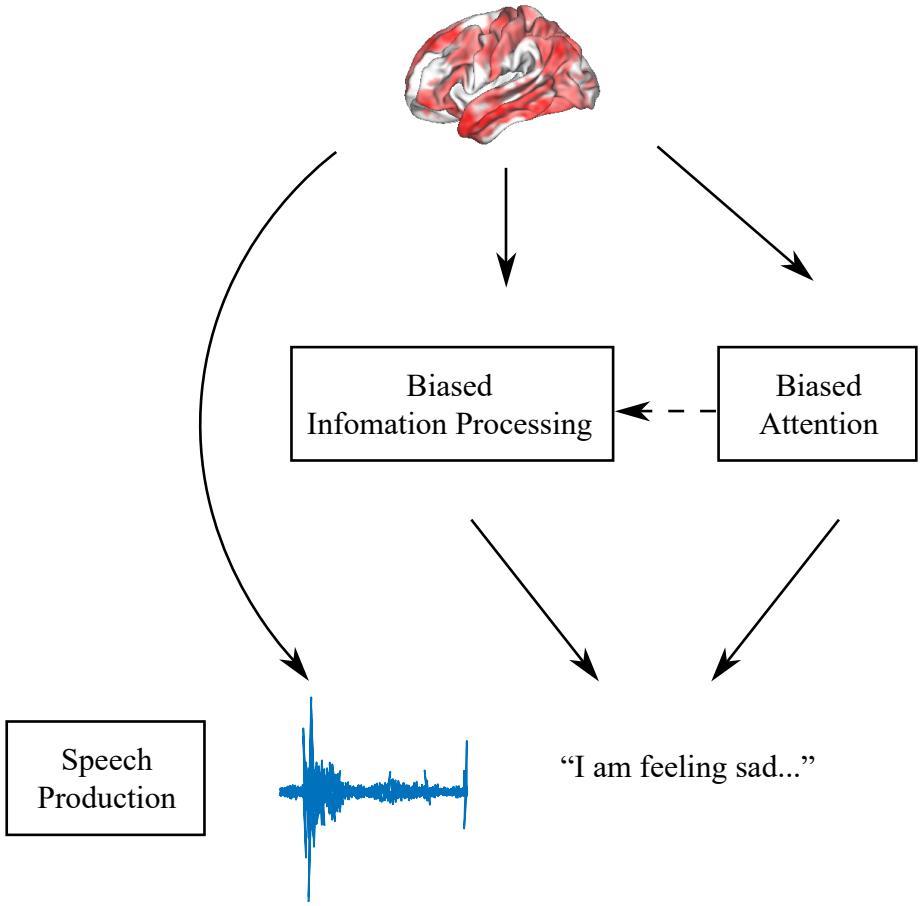

Figure 2.1: The figure shows that depression can influence speech in two primary ways. Firstly, alterations in the brain regions associated with language processing directly impact speech production. Secondly, alterations in other brain regions can also induce biased attention and information processing, which may trap the individuals in a state of persistent negativity. Consequently, this has an indirect effect on the content of their speech.

first delineate the process of voice generation, followed by an introduction to the changes in this process in depression. Then we will introduce speech variations observed in individuals with depression and this is the main motivation why we establish a new corpus including read and spontaneous speech.

2.2.1 Voice Generation

The control of motor muscles in voice generation includes the muscles of lungs, larynx and vocal tract. The process commences with the generation of pulmonary pressure by the lungs, followed by sound creation through phonation in the glottis of the larynx and finally, the modification of these sounds into distinct vowels and consonants by the vocal tract.

The vocal cords, located within the larynx and formed by two masses, are a critical component of the vocal tract. The inner edges of these masses create an opening known as the glottis. This structure plays an important role in voice generation, in which the larynx controls the vocal

cords to generate sound.

The lungs control the inhalation of air by expanding the surrounding rib cage and maintaining a steady flow of air during voice generation through precise muscle control. In the meanwhile, the vocal cords are opened and closed rhythmically, controlled by the larynx, to facilitate the passage of air through the glottis, thereby generating sound. Such a process introduces rapid vibrations of the vocal folds due to aerodynamic phenomena, leading to a sequence of vibratory cycles that produce sound. The larynx modulates the tension of the muscle, partly closing the vocal cords to control the size of the glottis and the airflow through it, thus altering the voice.

The fundamental frequency of the sound generated depends on the length, size, and tension of the vocal cords. For example, the average frequency for adult males, adult females and children are about 125 Hz, about 210 Hz, and over 300 Hz, respectively, mainly due to differences in larynx characteristics. Furthermore, the shape modification of the vocal tract is mainly determined by the articulators (e.g. tongue, palate, cheek, and lips) which modulate the sound emanating from the larynx to produce various phonemes.

Compared to any other human mechanical activity, more motor fibres are involved in the process of voice generation, in which around 100 muscles collaborate precisely in temporal respect, with the primary muscle groups involved being the laryngeal, respiratory, and articulatory muscles (Kent, 2000). In particular, the laryngeal muscles, including five muscle groups to control nine cartilages, collaborate with the diaphragm and intercostal muscle groups for respiration as well as those muscles controlling the lower jaw, tongue, lips, and velum for voice generation (Singh, 2019; Sataloff et al., 2007; Cummins, Scherer, et al., 2015).

From the perspective of acoustic theory, the source-filter model was proposed to analyse speech production and vocal acoustics (Singh, 2019; Almaghrabi et al., 2023). This model includes two stages, sound source modelling and conversion to speech signals. In the first stage, an excitation signal $x(t)$ often modelled by a periodic pulse train with spacing τ is used to represent sound signals generated from the glottis, while in the second stage, a filter with a continuous impulse response amplifies and attenuates the frequency response of the signal. Such a stage can represent resonant properties of the vocal tract and allow to define a transfer function $v(t)$ (Singh, 2019; Rabiner & Schafer, 2010; Almaghrabi et al., 2023). Therefore, the speech signal $s(t)$ can be defined as follows:

$$s(t) = x(t) * v(t), \tag{2.1}$$

where speech signal $s(t)$ is the result of convolving $x(t)$ with $v(t)$, and is periodic τ in the time-domain. When it comes to the frequency domain, such equation can be converted with the Fourier transform as follows:

$$S(j\omega) = X(j\omega) \cdot V(j\omega), \tag{2.2}$$

where $\cdot$ is the multiply operation.

2.2.2 Changed Voice Generation in Depression

As mentioned in the previous subsection, the process of voice generation involves many groups of muscles cooperating, however, many studies have reported that this process is affected by depression, leading to qualitative and quantitative changes in speech production (Kiss & Vicsi, 2017b; Cannizzaro et al., 2004). In particular, depressed patients exhibit psychomotor retardation that shows slowing of physical movements and thoughts (Cannizzaro et al., 2004). Such a behavior also reflects on the speech production that shows sluggishness and motor disorder in vocal articulation (Williamson et al., 2013). Furthermore, compared to healthy individuals, such an abnormal process leads to a longer response time, longer pause time and slower speech rate in depressed individuals (Yamamoto et al., 2020). This indicates that the speech observed in depressed individuals seems to be different from healthy people.

The above phenomena attracted the interest of researchers. According to Cummins, Scherer, et al. (2015), depressed patients have changes in muscle tension that affect the vocal tract and articulatory movement during voice production. For example, "[...] disturbances in laryngeal muscle tension will influence vocal fold behaviour whilst changes to respiratory muscles will affect subglottal pressure [...]" (Cummins, Scherer, et al., 2015). As a result, the prosody and quality of the speech are affected by the changed muscle tension and control. In particular, Y. Yang et al. (2012) revealed that the prosody features can reflect depression severity. From this perspective, the speech produced by articulators (in control of muscles in speech production) is expected to convey depression-related information.

In fact, cognitive impairments and psychomotor retardation can generate more speech errors (Cummins, Scherer, et al., 2015). For example, compared to healthy individuals, more phonation and articulation errors were observed during the speech of depressed patients (Christopher & MacDonald, 2005). Similarly, such a higher frequency of errors was also shown in depressed participants during a different speech task (Stasak et al., 2021). Furthermore, such speech errors can be magnified by external interference or in communication with others (Xie et al., 2019). Although speech disfluency is common in spontaneous speech (Johnson, 2004), a greater number of abnormal pauses and speech errors, resulting from abnormal cognitive-motor impact, were also observed in the simple reading tasks for depressed individuals (Tao et al., 2020; Stasak et al., 2021).

2.2.3 Speech Features in Depression

An important and frequently reported aspect of research in the field of depression-related speech is the investigation of alterations in speech features. Depression affects muscle tension and modifies salivary and mucus secretion (Cummins, Scherer, et al., 2015), thus leading to changes in speech features. Speech features employed for depression detection can be categorised into four groups, namely *source features*, *spectral features*, *prosodic features*, and *formants fea-*

tures (D. M. Low et al., 2020). There is a simple and standard speech feature set proposed in Schuller et al. (2009), which includes these four aspects of features.

Source Features

Source features provide information related to glottis when producing the voice (Cummins, Scherer, et al., 2015). Such features play an important role in characterising voice quality, which refers to the changes, unrelated to prosodic elements such as pitch, in the vibrations of vocal cords and the shape of vocal tract as perceived auditorily (Cummins, Scherer, et al., 2015). According to D. M. Low et al. (2020), voice quality features and glottal features, as two main types of source features, were suggested to model the vibration of vocal folds and the glottal flow during voice production. In general, the measurements of voice quality in depression detection include *jitter*, *shimmer* and *harmonic-to-noise ratio* (HNR).

To this end, features measuring voice quality (e.g., jitter, shimmer) have been proposed to describe small cycle-to-cycle changes in glottal pulse, and then showed a strong connection to depression (Quatieri & Malyska, 2012). Similar to jitter and shimmer, HNR describes a ratio of harmonics to inharmonic components, which are higher in depressed patients due to the altered patterns of airflow in the speech production (L.-S. A. Low et al., 2010).

- Jitter is impacted by the control of vocal fold vibration, and it measures the variation of successive glottal periods, typically ranging between 0.50% and 1.00% in adults (Teixeira et al., 2013). Jitter is defined as follows:

$$Jitter = \frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i-1}|, \tag{2.3}$$

  where N is the total number of extracted periods and T_i is the time period of the glottal pulse.

- Shimmer reflects the peak values of a signal and it is affected by glottal resistance or mass lesions on the vocal cords (Teixeira et al., 2013), typically ranging between 0.05 dB and 0.22 dB (Haji et al., 1986). Shimmer measures the amplitude of the variation of successive glottal periods, defined as follows:

$$Shimmer(dB) = \frac{1}{N-1}\sum_{i=1}^{N-1}\left|20 * log\left(\frac{A_{i+1}}{A_i}\right)\right|, \tag{2.4}$$

  where N is the total number of extracted periods, A_i is the amplitude measured from peak-to-peak.

- HNR measures the ratio between periodic and non-periodic components (harmonics to

inharmonic components) in the voice. HNR is defined as follows:

$$\text{HNR} = 10 \cdot \log \frac{\text{ACF}(T_0)}{\text{ACF}(0) - \text{ACF}(T_0)}, \tag{2.5}$$

where T_0 is the fundamental period determined by mathematical fundamentals presented by Boersma et al. (1993), and ACF is the autocorrelation function.

Given that depression-related vocal cord vibration (Ozdas et al., 2004) results from the changes in subglottal pressure and vocal cord tension (Sundberg et al., 2011), Quatieri & Malyska (2012) found that jitter and shimmer show a strong connection to depression. In particular, Silva et al. (2021) found that depressed patients tend to have significantly higher values of jitter and shimmer compared to healthy individuals. Such results reflect the glottal irregularities in phonation in depressed individuals, affecting the dynamic coordination of the larynx motor. In addition, HNR is also correlated with depression severity significantly (Quatieri & Malyska, 2012) and is higher in depressed patients due to the altered patterns of airflow in the speech production (L.-S. A. Low et al., 2010). These results also reflect turbulent glottis led by muscle incoordination. The above source features showing differences in depressed individuals were investigated in Table 2.1.

Prosodic Features

Prosodic features describe the variations in the perceived acoustic properties, such as rhythm, stress, and intonation of speech, reflecting the manner of speaking (Cummins, Scherer, et al., 2015). This type of feature includes speaking rate, pitch (F0, the fundamental frequency) and loudness.

Pitch is a perceptual property of speech and it represents the glottal excitation rate (Sethu et al., 2009). Quantifiable as a frequency such as F0 (fundamental frequency), pitch corresponds to the lowest frequency of a periodic vocal cord vibration. The variation in F0 values is typically affected by gender and mainly depends on vocal cord structures, including anatomy and larynx size (Benesty et al., 2008). According to paralinguistic investigations on depression speech, researchers found depressed patients tend to show speech abnormalities in prosodics including decreased pitch and pitch range (Cummins, Scherer, et al., 2015). In particular, Hussenbocus et al. (2015) observed that males with depression tend to have decreased F0 values while females with depression tend to have increased F0 values. Despite these findings, some studies reported no correlation between F0 and depression (Mundt et al., 2012; Quatieri & Malyska, 2012; Y. Yang et al., 2012). Such discrepancy may arise from inconsistent ground truth (clinical diagnosis vs self-assessments) or variations in speech content (reading vs interview) across different studies.

The perception of loudness is associated with the intensity of the speech signal and it typically depends on glottal excitation strength and sub-glottal pressure etc (Seshadri & Yegna-

narayana, 2009). The previous study suggested that the decrease in loudness is associated with depression (K. R. Scherer, 2013). However, this relationship has been contested in some studies (Cummins, Scherer, et al., 2015). For example, Alpert et al. (2001) found that depressed individuals spoke louder than non-depressed individuals. In recent years, researchers typically use the root mean square of the energy instead of loudness to measure the energy in speech, see Section 3.3.1 for more details.

Given the effects of psychomotor retardation among depressed patients, researchers explored timing-based features in order to capture these delayed articulatory movements. In this perspective, features such as prosodic timing measures (Mundt et al., 2012), speech rate and phoneme rate (Trevino et al., 2011) have been proposed. In particular, many studies report depressed patients to have a slower speech rate compared with controls (Cummins, Scherer, et al., 2015; Hönig et al., 2014; Alghowinem et al., 2012). (Trevino et al., 2011) also explains depressed speech rate from the phoneme level of speech production. One possible explanation for such a situation is that depressed patients tend to have more pauses during speaking (Mundt et al., 2012; Liu, Li, et al., 2017), thus taking more time in speech. Furthermore, evidence suggests that the duration of pauses in speech from depressed individuals is typically longer than those from non-depressed individuals (Tao et al., 2020). The prosodic features that have been identified as showing significant differences in studies related to depression are presented in Table 2.1.

Other prosodic features such as zero-crossing rate (ZCR) and voicing probability (VP) are also important and have been revealed differences in depressed people, see Section 3.3.1 for more details.

Formants Features

Formant features, often called filter features, represent frequency peaks in the spectral distribution where the speech signal concentrates a high proportion of acoustic energy (Abhang et al., 2016). These features are relevant to the characteristics of the vocal tract and are influenced by factors such as muscular tension and the production of salivation or mucus (Cummins, Scherer, et al., 2015). Formant features capture the information related to the acoustic resonances of the vocal tract, thereby reflecting the coordination of speech articulators (Williamson et al., 2013). Despite a signal theoretically being capable of containing an infinite number of formants, only a limited number fall within the range of human auditory perception (Özseven & Düğenci, 2018). The top two formant frequencies, commonly referred to as F1 and F2, are critical in specifying the quality of a vowel (Benesty et al., 2008). The subsequent three formant frequencies (F3, F4, F5) define the property of the voice (Stanek & Polak, 2013).

The reduction of articulatory effort in depressed patients resulting from psychomotor disturbances, thus has an impact on the formant features (Flint et al., 1993). In general, depressed patients exhibit a decrease in average formant frequencies, such as F1 and F2 (Kiss & Vicsi, 2017b; Flint et al., 1993; Vicsi et al., 2012). However, such observations are not consistently

Table 2.1: The survey of widely-used speech features in terms of source features, prosodic features, formants features and spectral features shows significant differences between depressed speech and healthy speech.

Features types	Features	Studies
Source Features	Jitter&shimmer	(Ozdas et al., 2004; Carding et al., 2009; Quatieri & Malyska, 2012; Alghowinem, Goecke, Wagner, Epps, Gedeon, et al., 2013; Sahu & Espy-Wilson, 2016; Silva et al., 2021)
	HNR	(Quatieri & Malyska, 2012; Horwitz et al., 2013; Jia et al., 2019, 2020; Albuquerque et al., 2021)
Prosodic Features	F0	(Alghowinem, Goecke, Wagner, Epps, Gedeon, et al., 2013; K. R. Scherer, 2013; Guidi et al., 2015; Hussenbocus et al., 2015; Simantiraki et al., 2017)
	Loudness	(Alghowinem, Goecke, Wagner, Epps, Gedeon, et al., 2013; J. Wang et al., 2019)
	ZCR & VP	(Alghowinem, Goecke, Wagner, Epps, Gedeon, et al., 2013; Morales & Levitan, 2016)
	Pauses & speech rate	(Alpert et al., 2001; Cannizzaro et al., 2004; Mundt et al., 2012; Morales & Levitan, 2016; Liu, Kang, et al., 2017)
Formants Features	F1/F2/F3	(France et al., 2000; Moore II et al., 2007; Vicsi et al., 2012; Shannon et al., 2016; Kiss & Vicsi, 2017a,b)
	Vocal tract coordination	(Williamson et al., 2013, 2014, 2016, 2019; Tao, Ma, et al., 2023)
Spectral Features	MFCCs	(Yingthawornsuk et al., 2006; Quatieri & Malyska, 2012; Cummins, Epps, & Ambikairajah, 2013; Kiss & Vicsi, 2017b; Taguchi et al., 2018; J. Wang et al., 2019)
	Mel-filtered band energies	(Kiss & Vicsi, 2017a)

reported across all studies. For example, Mundt et al. (2007) reported that there was no significant correlation between F1 variability and depression. Similarly, France et al. (2000) observed increases in the locations of formant frequencies (F1-F3). The potential source of this

inconsistency could lie in the complex relationship between the dynamics of the source and the filter (Cummins, Scherer, et al., 2015). Another potential explanation for such a discrepancy may be the different antidepressant medications used across studies. Some of these medications may have anticholinergic effects that decrease saliva production, thereby leading to a drier vocal tract and mouth, and therefore, this could subsequently interfere with the formants and the distribution of energy (France et al., 2000; Cummins, Scherer, et al., 2015).

Psychomotor retardation can result in vocal tract constriction and/or a deficit in motor coordination, which in turn may affect the formant features in individuals with depression (Quatieri & Malyska, 2012; Flint et al., 1993). Drawing on observations of psychomotor retardation in depressed patients, Williamson et al. (2013) proposed more sophisticated features related to the coordination of vocal tract articulation across variable time scales. Specifically, Williamson et al. (2013) focused on changes in the correlation across different time scales of formant frequencies, thereby reflecting changes in formant features of depressed individuals. The reduction of vowel space among depressed patients corroborates this, indicating less articulate speech and altered formant features (S. Scherer et al., 2015). Additionally, a decreased acoustic feature space in depressed patients was also reported (Cummins, Sethu, et al., 2015). Therefore, the changes in formant features provide a direction for investigating depressed speech from an articulatory perspective, and thus allow researchers to detect depressed speakers from irregular articulatory information (Huang et al., 2019b,a). Table 2.1 presents formant features investigated in the studies related to depressed speech.

Spectral Features

Spectral features, representing the distribution of frequencies in human speech signals over a specific duration, can be analyzed for their individual frequency components. They are capable of capturing variations in prosodic, phonetic, and articulatory information that occur due to speech motor control (Almaghrabi et al., 2023). These features have been utilized to reflect the psychological state of the speaker (Cummins, Scherer, et al., 2015). In addition, spectral features were also observed to represent the relationship between the movements of articulators and the shape of vocal tract (W. Pan et al., 2019). Therefore, spectral features are also expected to capture psychomotor retardation and disturbances in muscle tension (Almaghrabi et al., 2023). Table 2.1 presents spectral features showing significant differences between depressed and non-depressed individuals.

In the studies investigating the speech of depression, it has been observed that a relative shift in spectral energy is common among individuals affected by the pathology. This shift typically involves the spectral energy transitioning from lower to higher frequency bands (France et al., 2000; Ozdas et al., 2004). Specifically, the averaged energy values in low-frequency bands (65–400 Hz) tend to shift towards higher values, and those in higher frequency bands (1330–5735 Hz) lean towards lower ones (Kiss & Vicsi, 2017b). This energy shift may be an

outcome of the increased tension in the vocal tract or vocal cords, thus altering properties in the resonance of the vocal tract (Cummins, Scherer, et al., 2015; K. R. Scherer, 1986).

Among the spectral features, mel-frequency cepstral coefficients (MFCCs) are commonly used. MFCCs are obtained by taking logs of powers on a nonlinear mel scale of frequency and transforming them by a linear cosine. The main advantage of using MFCCs is that the frequency bands on the mel scale are equidistant in perception, therefore, compared to using the spectrum with linearly-spaced frequency bands, it mimics the response of the human auditory system more accurately. Studies have demonstrated that MFCCs, representing the speech spectrum of a speech segment, show decreased temporal variations with increasing severity of depression (Cummins, Epps, Sethu, et al., 2013). For example, MFCC5 and MFCC7 values are found to be lower in depressed speech, which can be leveraged as effective indicators for depression identification through voice analysis J. Wang et al. (2019). Furthermore, a significant increase in MFCC2 was observed in the speech of depressed patients, such a difference that remains unaffected by factors such as age or gender (Taguchi et al., 2018). These variations in MFCCs between depressed and non-depressed speech reflect the differences in vocal tract configuration from an anatomical perspective (Zhu et al., 2012).

The standard procedure in depression detection involves the concatenation of MFCCs with their time derivatives (i.e., delta), which captures the temporal variation of MFCCs and provides additional information relevant to depression. For instance, MFCCs combined with time derivatives were found to have a significantly negative correlation with the severity of depression (Cummins, Epps, Sethu, et al., 2013). These findings indicate the distinct spectral characteristics of depressed speech, showing the potential of such features used for automatic depression detection, see Section 2.3.

2.2.4 Changes of Depressed Speech in Interview

Complementing the findings on acoustic errors, there are some speech errors of depressed speakers in the interview. Table 2.2 presents related tasks used in the studies investigating spontaneous speech. For example, Rubino et al. (2011) found that, compared to healthy individuals, depressed individuals tended to have more referential failures which are errors replacing words (i.e., malapropisms) in spontaneous speech. According to the definition of malapropism, it is a type of error involving the substitution of an intended word with an incorrect one. This misused word is typically unrelated in meaning but shows some similarities, such as in word pronunciation, syllable length, or grammatical category (Fay & Cutler, 1977). In particular, given that depressed patients exhibited passive avoidance strategies (Holahan & Moos, 1987; Holahan et al., 2005), it is expected that, compared to non-depressed individuals, depressed speakers may be less inclined to self-correct their speech errors.

Another study comparing the speech patterns of depressed and non-depressed individuals found that depressed individuals are more likely to express negative emotions when describing

Table 2.2: The survey of tasks used for investigating the spontaneous speech of depressed individuals.

Studies	Participants	Tasks
(Rude et al., 2004)	124	Writing about feeling in the college
(Molendijk et al., 2010)	412	Writing about their life
(Fast & Funder, 2010)	181	Interviewed by clinician
(Jarrold et al., 2011)	26	Structured interviews
(Zimmermann et al., 2013)	118	Semi-structured interview
(Van der Zanden et al., 2014)	234	Answering to a set of questions
(Bernard et al., 2016)	136	Writing about feeling in the college
(Nook et al., 2017)	107	Writing responses to neutral images

their feelings (Long et al., 2017). Moreover, a linguistics meta-analysis further confirmed these findings, indicating that depressed individuals tend to use fewer positive words in their speech compared to healthy individuals (Tølbøll, 2019). This observation aligns with Beck's cognitive model, confirmed by the study from a perspective of spontaneous speech (Rude et al., 2004). In particular, depression has an impact on the ways in which individuals perceive, interpret, and communicate their experiences by the cognitive mechanisms that promote increased self-focus and negative thinking (Bernard et al., 2016). In particular, depressed individuals incline to use more first-person pronouns, and such a finding reflects self-focus in depression (Tackman et al., 2019). The meta-analysis further confirmed this phenomenon, and suggested that the use of first-person singular pronouns significantly correlated with depression, importantly, such correlation was not biased (Holtzman et al., 2017).

2.3 Overview of Automatic Depression Detection

In the last section, we introduced the various differences in terms of read and interview speech that result from changes in the vocal tract caused by depression. This is the main motivation for this thesis why to establish the Android Corpus, collecting read and spontaneous speech. These differences in such speech provide a reliable basis for developing techniques for depression detection. In this section, we will review studies using speech for automatic depression detection.

2.3.1 Advantages of Automatic Depression Detection

The first advantage of such a technique is that the outcome may be robust and independent to the language used in the system. For instance, Alghowinem et al. (2016) investigated the use of German and English recordings from clinical interviews for depression detection. In particular, such speech recordings had different accents, such as Australian and American. The results suggested that the application of SVM models to speech recordings in the different language yielded comparable classification performance, with an average recall of 89.5%. Even when all speech recordings were combined, depression detection remained effective, achieving an average recall of 74.6%. In the similar vein, Kiss et al. (2016) investigated the acoustic features extracted from read speech of Hungarian speakers (54 patients, 73 controls) and Italian speakers (11 patients and 11 controls). The results suggested that the variance in acoustic features between healthy and depressed speech was found to be similar in both languages, and there was no significant difference in the classification outcomes when using different languages. The above findings indicate that healthy and depressed speech carries comparable separation tendencies across different languages.

Another advantage in the use of automatic depression detection is that it allows to be applied to large-scale depression screening. For example, computer-automated methods can be employed for telephone data collection (Mundt et al., 2007, 2012). A voice-controlled agent, allowing for hands-free interactions, can be utilized for remote mental health monitoring, thus overcoming challenges such as cost of health care and time, stigma, and accessibility that often hinder mental health care access (D. M. Low et al., 2020). Furthermore, this technique can comply with standard protection of sensitive personal data. For example, healthcare organizations have started to analyze electronic health data with devices that meet standard protection for patients data (Brewer et al., 2022). As a result, The trend towards developing privacy-preserving techniques for automatic depression detection is on the rise (X. Xu, Peng, et al., 2021; Almaghrabi et al., 2023; Koops et al., 2023).

2.3.2 Read and Spontaneous Speech of Depression

In recent years, the computing community has made substantial efforts towards automatic depression detection with speech (Cummins, Scherer, et al., 2015; Wu et al., 2022). After observing the difference in speech production between depressed patients and healthy individuals (Cummins, Scherer, et al., 2015), researchers have tried to identify which type of speech is best suited to maximize the difference in speech. The first type is spontaneous speech, mainly because it is the closest format to the clinical depression diagnosis. In clinical settings, depressed patients are typically diagnosed by answering a series of questions in questionnaires that reflect psychometric properties, such as *Hamilton Depression Rating Scale*. Another reason is that spontaneous speech is considered to be more capable of expressing the emotional state of

the speaker (Mitra & Shriberg, 2015).

In this respect, many studies have started to investigate the effectiveness of spontaneous speech in automatic depression detection because it provides more acoustic variability (Alghowinem, Goecke, Wagner, Epps, Breakspear, & Parker, 2013; Jiang et al., 2017; Alghowinem et al., 2012; Aloshban et al., 2020, 2022; Alsarrani et al., 2022). For example, tempo-related features such as pause lengths and articulation rate are useful in spontaneous speech (Kiss & Vicsi, 2017a). This is in line with the phenomenon that depressed patients tend to speak with significantly longer pauses than non-depressed speakers (Alghowinem, Goecke, Wagner, Epps, Breakspear, & Parker, 2013; Esposito et al., 2016). Additionally, depressed patients tend to have more referential failures, such as word replacement errors where an incorrect substitution word with an unrelated meaning but a similar pronunciation is used instead of the intended word (Rubino et al., 2011). However, such an observation seems to be affected by the lower participation of depressed patients when speaking spontaneously, during which they may typically attempt a few times before speaking up or keeping silent (He et al., 2022).

The aforementioned issue can be addressed by providing predefined text, i.e., read speech, because it requires less participation from depressed patients. On the one hand, the unimportant nature of the material being read limits the impact of variability related to speech content, and this provides researchers with a perspective to investigate acoustic properties of speech. For example, formants trajectories are more effective in read speech due to their sensitivity to the content of speech (Mitra & Shriberg, 2015; Kiss & Vicsi, 2017a). On the other hand, the predefined text makes the speakers more careful with the articulation of speech, making it easier to capture acoustic phonetic information (Mitra & Shriberg, 2015). In addition, read speech provides specific tempo-related features, such as read speed, indicating that depressed patients may tend to read slower (Tao et al., 2020), which is in line with the phenomenon of psychomotor retardation among depressed patients (Williamson et al., 2013). However, the disadvantage of read speech is that it suppresses the emotional state of the speaker (Mitra & Shriberg, 2015).

Researchers have found some acoustic features (e.g. variability of fundamental frequency and mel-frequency cepstral coefficients) exhibit common variability in the read and spontaneous speech (Long et al., 2017), but they differ from each other phonetically and phonologically (Kiss & Vicsi, 2017a). Recently, more and more studies have taken both of them into account for depression detection (Jiang et al., 2018; Lu et al., 2021; Rejaibi et al., 2022; Du et al., 2023). However, the main issue is that different types of speech are fed into the model in the same way, which leads to the loss of unique information related to each type. The approach proposed in this work addresses the problem by jointly considering the specific information of read and spontaneous speech.

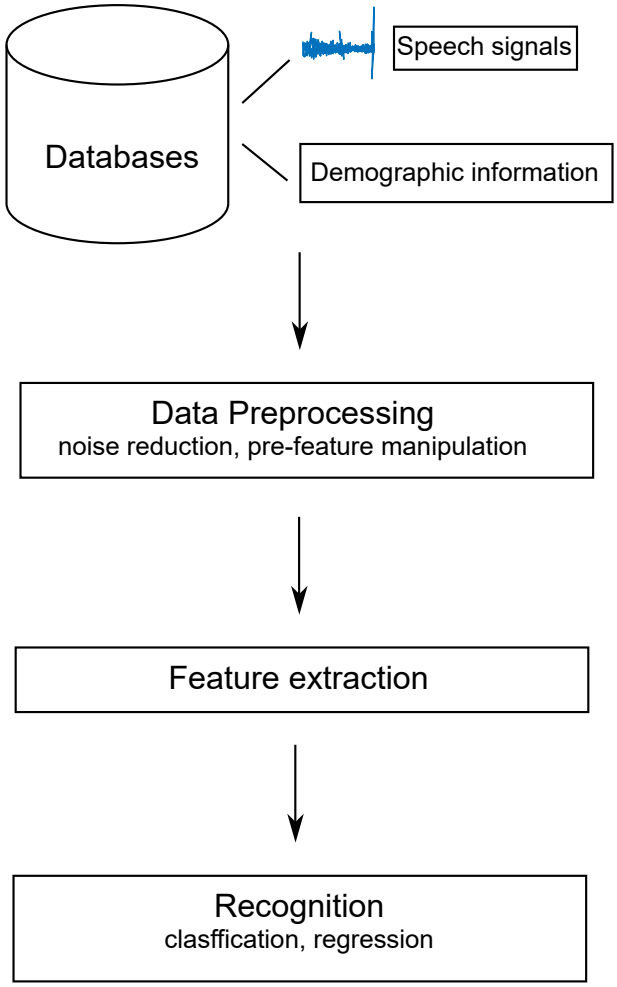

Figure 2.2: The overview of steps in the automatic depression detection system.

2.3.3 Automatic Depression Detection Research Pipeline

Figure 2.2 shows the overall framework commonly used in automatic depression detection. This framework typically includes four steps, namely *data acquisition*, *data preprocessing*, *feature extraction* and *depression recognition*.

Data Acquisition

The data of participants including speech signals and demographic information are stored in databases, the Andriods Corpus, see Chapter 3 for more details. The speech signals typically are the recordings during the tasks performed by participants, such as reading aloud a passage, describing pictures, communicating with others etc. The demographic information might be of help in detecting depressed speakers and includes age, gender, education level etc. In addition to the speech signals, other types of data used in the automatic depression detection system include electroencephalogram (EEG) J. Shen et al. (2020), functional near-infrared spectroscopy (fNIRS) (R. Wang et al., 2021), accelerometers (sensors) (Burton et al., 2013), Global Position-

ing System (GPS) (X. Xu, Chikersal, et al., 2021) and WiFi (Yue et al., 2020). In this thesis, we only survey automatic depression detection using speech signals.

Data Preprocessing

Depending on the task (classification or regression), the data of all participants should be divided into two groups according to the ground truth, a group of depressed patients and a group of healthy controls. For a classification task, such as detecting a depressed speaker, the labels for the depressed patients are determined by the ground truth (Cummins, Scherer, et al., 2015; Kiss & Vicsi, 2017a; Aloshban et al., 2020; Tao, Ge, et al., 2023). However, for a regression task, such as predicting the severity of depression in patients, the labels depend on questionnaire scores, which serve to indicate depression severity (Cummins, Scherer, et al., 2015; Williamson et al., 2019; Dumpala et al., 2021; He & Cao, 2018).

The main steps of data preprocessing include noise reduction and pre-feature manipulation. Noise reduction is the process of removing noise and outliers from speech signals. pre-feature manipulation is normalizing, filtering or segmentation to redistribute the features in the raw data for the next step of feature extraction.

Normalization is crucial data manipulation techniques, transforming the value range to a 0-1 scale and modifying the influence of specific data points, respectively. These processes enhance the data quality and integrity, thereby avoiding anomalies and inconsistencies. Z-score normalization and min–max normalization are the most widely used normalization methods (Patro & Sahu, 2015). The former transforms the distribution of probability into a standard Z distribution with an area of 1, while the latter converts the minimum value in the dataset into 0 and the maximum value into 1. These transformations both convert values into a constrained range, enabling an improved understanding of the relative impact of various data points.

Segmentation, the process of dividing data into distinct segments, varies based on data types. In speech signal processing, segmentation can be used to identify boundaries between words, syllables, or phonemes, facilitating linguistic analysis of spoken natural languages (Schuller & Batliner, 2013). Furthermore, segmentation can isolate irrelevant speech data, such as the voice of interviewer during communication (Aloshban et al., 2021) or silence before the start of experiments (Tao et al., 2020). It can also be employed for detecting speaking through Noise-Robust segmentation (Khorram et al., 2016), voice activity detection(Karam et al., 2014), and speech recognition(Mattys et al., 2012).

Pre-feature manipulation involves altering the data into a more useful form in order to be in line with the objective, without necessarily removing any information before it is input into the model. An example of these methods is signal transformation, which mathematically converts the data into a different form for specific analysis, such as the Fast Fourier Transform (FFT) that converts time-domain data into frequency-domain data (Brigham, 1988).

Both filtering and segmentation help exclude noise data, and this allows the extraction of

more accurate speech features for subsequent analysis. However, the removal of some data might inadvertently lose important information from the signal, or distort the original signal. As a result, this might lead to over-fitting during the recognition stage, making it challenging to use in some environments with high noise levels.

Feature Extraction

The process of feature extraction converts speech signals into feature vectors. These vectors aim to encapsulate the maximum amount of information present in speech signals, while minimizing redundancies (Kinnunen & Li, 2010). The choice of extracted features depends on the specific goals of the research and varies among studies, but commonly includes linguistic and acoustic features, as discussed in Sections 2.1 and 2.2. Within the context of depression detection, the extracted features proposed in the Audio/Visual Emotion Challenge (AVEC) 2009 has seen widespread use for inferring human emotional and psychological states (Schuller et al., 2009). During extracting features, a commonly used method involves segmenting speech signals into multiple short windows, then extracting predefined speech features from each window. This method allows for a more precise representation of syllables or phonemes within spoken words.

Another feature extraction approach has been proposed for depression detection is the use of i-vectors (Cummins et al., 2014). I-vectors have been extensively proposed for speech recognition applications (Dehak et al., 2010). Frame-level features are acquired by adapting the Universal Background Model, which characterizes the feature distribution of acoustic space, to a series of specific speech frames, thereby estimating the parameters of utterance-dependent Gaussian Mixture Models (GMM). The i-vector space represents a low-dimensional subspace where supervectors represented by GMM are linearly transformed, retaining as much variability or valuable information in the raw vector space. The i-vector extraction can be expressed as:

$$\Phi_x = \Phi + T\omega, \tag{2.6}$$

where Φ_x is the supervectors corresponding to the speech utterance, Φ is the supervectors corresponding to the UBM, T is a low-rank matrix capturing all relevant variability and ω is the i-vectors (Kenny et al., 2005). However, the disadvantage of this approach is that the extraction process of i-vectors depends on the assumption that the extracted features follow a Gaussian distribution. However, such an assumption may not be suitable for all types of speech data, leading to inaccurate estimation of i-vectors.

An alternative feature extraction approach for depression detection (Egas-López et al., 2022) involves the use of x-vectors, which are also commonly used for speech recognition (Snyder et al., 2017, 2018). Conceptualized as a neural network-based feature extraction method, the x-vector approach generates fixed-dimensional embeddings to represent unfixed-length utterances. The extraction of x-vectors includes following steps. Firstly, the structure of the Deep Neural

Network (DNN) is such that the frame-level layers with time-delay architecture. The output from the previous frame-level layer is concatenated and as input fed to the current layer. From this perspective, this approach takes the temporal context information into account. Next, the stats pooling aggregates output from the last frame-level layer and calculates the average and standard deviation. These are then concatenated and fed to subsequent segment layers to extract the x-vectors embeddings. The advantage of this approach is that it captures speaker-specific information throughout the entire speech signal, and thereby resulting in a fixed-size vector that is irrelevant the utterance length. However, the disadvantage is that the x-vectors are sensitive to the choice of layer used for extraction, and such a choice is empirical and varied across speech data.

Depression Recognition

In this step, the extracted feature vectors are fed to the model for recognition. These models are typically developed based on machine learning or deep learning algorithms. In the next section, we will introduce some algorithms that are widely used in the field of depression detection, such as Support Vector Machine (SVM), Long-short Term Memory neural networks (LSTMs). Table 2.3 presents a survey of algorithms used for depression detection with corresponding experiments and results, evaluated by standard evaluation metrics, see 2.3.6 for more details.

2.3.4 Algorithms in Depression Detection

Support Vector Machine

A Support Vector Machine (SVM) is a powerful and simple algorithm and it has been widely utilized in various domains, including speech and vision (Zeng et al., 2007), particularly for classification tasks. The SVM operates under the principles of supervised learning, where it constructs and compares multiple hyperplanes with the goal of identifying the one that achieves the maximum separation between data points from given classes (Larsen et al., 2015). In other words, SVM aims to discriminate categories with a gap of maximal width. The data points are projected into this space and categorized based on their location relative to the gap (Cristianini & Shawe-Taylor, 2000; Schölkopf & Smola, 2002).

The SVM algorithm can be mathematically formulated as follows:

$$\min_{\omega,b,\varepsilon_i} \frac{1}{2}\|\omega\|^2 + \lambda \sum_{i=1}^{m} \varepsilon_i \tag{2.7}$$

$$\text{s.t. } y_i(\omega^T \cdot x_i + b) \geq 1 - \varepsilon_i, \varepsilon_i \geq 0, i = 1,2,\cdots,m \tag{2.8}$$

where ω is the vector orthogonal to the hyperplane, λ is a constant to punish training errors, controlling the trade-off between maximizing the margin and minimizing errors, and b is a bias

Table 2.3: The survey of automatic depression detection studies. Abbreviations: PT - Depressed patients, HC - Healthy controls, Acc - Accuracy, Pre. - Precision, Rec. - Recall, F1. - F1-score, MAE - mean absolute error, RMSE - root mean squared error. The scores are averaged for the studies with multiple tasks.

Algorithms	Studies	Samples	Metrics	Scores	Experiments
SVM	(Kiss & Vicsi, 2017a)	48PT 25HC	Acc.	84.5%	Reading; Interview
	(Jiang et al., 2017)	85PT 85HC	Acc.	78.13%	Reading; Interview; Picture description
	(Long et al., 2017)	37PT 37HC	Acc.	78.02%	Reading; Interview; Picture description
	(Stolar et al., 2018)	29PT 34HC	Acc.	76.35%	Interview
	(Tao et al., 2020)	56PT 54HC	Acc.	84.5%	Reading
RF	(Espinola et al., 2021)	22PT 11HC	Acc.	87.55%	Interview
	(Stasak et al., 2022)	44PT 108HC	F1.	69.2%	"pataka" task
	(Wanderley Espinola et al., 2022)	28PT 12HC	Acc.	75.27%	Interview
MLP	(Dong & Yang, 2021)	292 participants	F1.	81.6%	Reading; Interview
	(Sun et al., 2022)	-	MAE	0.77	Audio and video tasks
CNN	(Saidi et al., 2020)	189 participants	F1.	69.0%	Interview
	(Y. Wang et al., 2022)	25PT 25HC	Acc.	77%	Reading; Interview; Picture description
	(Q. Wang & Liu, 2022)	76PT 81HC	F1.	82.4%	Reading
LSTM	(Z. Zhao et al., 2019)	189 participants	MAE	4.20%	Interview
	(Z. Zhao et al., 2020)	189 participants	MAE	4.28%	Interview
	(Y. Zhao, Xie, et al., 2021)	-	Acc.	90.2%	Interview
	(Rejaibi et al., 2022)	292 participants	RMSE	0.168%	Reading; Interview
	(Tao, Ge, et al., 2023)	55PT 54HC	F1.	88.0%	Reading

term.

In addition to classification tasks, SVM can be adapted for regression tasks (Support Vector

Regression), using the hyperplane itself. The main advantage of using SVM is that it provides good generalization performance (Schuller et al., 2011). As a result, it has been widely employed for automatic depression detection (Cummins, Scherer, et al., 2015; Kiss & Vicsi, 2017a; Long et al., 2017; Z. Pan et al., 2018; Scibelli et al., 2018; Tao et al., 2020). For example, Long et al. (2017) fed a set of extracted acoustic features into an SVM classifier, achieving 78.02% accuracy in depression detection. Alghowinem, Goecke, Wagner, Epps, Gedeon, et al. (2013) utilized recordings from 30 depressed and 30 non-depressed speakers for speech features extraction, and they reported an improved performance in recall with the average of 81.61% when using SVM and Gaussian Mixture Models jointly.

Furthermore, the use of SVMs in both read speech and spontaneous speech was effective in distinguishing between depressed and non-depressed speakers, with 86% and 83% classification accuracies, respectively (Kiss & Vicsi, 2017a). Another utilization of SVM in the field of depression detection involves evaluating the effectiveness of acoustic features in recognizing depressed speakers, for example, Tao et al. (2020) reported a 16.3% improvement in classification accuracy by incorporating pause features into the SVM model.

Random Forest

The Random Forest (RF) algorithm, as an ensemble learning method, is also widely used in the field of speech-based depression detection. This algorithm assembles a large number of decision trees during the training phase. Similar to the SVM, RF can execute both classifications and regressions. When applied to classification tasks, the result of RF depends on the majority output from the collection of trees. Conversely, for regression tasks, the result is determined by the mean predictive value returned by the individual trees.

A study conducted by Stasak et al. (2022) evaluated the classification performance by using acoustic features extracted from the speech signals, such as spectral, prosodic, landmark, and voice quality. The experiment involved 152 speakers, comprising both 108 controls and 44 depressed patients. The findings indicated a 10% increase in classification accuracy by the RF classifier. In a similar vein, Wanderley Espinola et al. (2022) constructed 300 trees to extract 33 speech features from the recordings of 78 speakers, achieving a classification accuracy of 75.27%. Moreover, McGinnis et al. (2019) utilized recordings from 71 children recorded during a three-minute speaking task, and found that the performance across various machine learning models including logistic regression, SVM, and RF are comparable in depression detection.

One of the key strengths of RF is its ability to lower variance and produce robust predictive outcomes due to ensemble averaging. However, combining diverse trees may introduce bias into the model. Given that RF works by generating a plethora of decision trees, it can be demanding in terms of memory, and therefore, may not be the most efficient model for data with a high volume of features.

Multi-Layer Perceptron

The Multilayer Perceptron (MLP), a type of feedforward artificial neural network (ANN) that is fully interconnected, has been widely used in various applications, such as pattern recognition (Abiodun et al., 2019) and emotion recognition (Basu et al., 2017). An MLP typically features a minimum of three layers: an input layer and an output layer at both ends, and one or more hidden layers between them. Each layer contains nodes (perceptions) that are interconnected with the nodes in the subsequent layer. Besides the input nodes, each node embodies a neuron utilizing a nonlinear activation function. In several MLPs, specific neurons utilize a nonlinear activation function to emulate the firing frequency of action potentials of biological neurons. The widely used activation functions are sigmoid function $\sigma(x)$, tangent function $tanh(x)$ and ReLU function $f(x)$, defined as follows:

$$\sigma(x) = \frac{1}{1 - e^{-x}} \tag{2.9}$$

where x is the output from a neuron. Similarly, the tangent function $tanh(x)$ is defined as follows:

$$tanh(x) = \frac{e^{2x} - 1}{e^{2x} + 1} \tag{2.10}$$

The ReLU function $f(x)$ is defined as follows:

$$f(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0 \end{cases} \tag{2.11}$$

Owing to their ability to model intricate relationships between inputs and outputs and to discover patterns within data, MLP networks are frequently employed in research for detecting depression (Alghowinem, Goecke, Wagner, Epps, Gedeon, et al., 2013; Alsarrani et al., 2022; Sun et al., 2022; Du et al., 2022). For example, Du et al. (2022) reported a better performance in the use of MLP for depression detection when compared to other traditional models, achieving an average accuracy of 78.15% using the same dataset. Furthermore, Sun et al. (2022) extended the use of MLP to accommodate multimodal data for depression detection, resulting in an improved F1-score of 85.5%.

However, the architecture of the network is typically determined by the number of perceptrons for each layer, the number of hidden layers used, activation function used after each layer etc, which is nontrivial and adds complexity to the model. As a result, MLP tends to be overfitting and needs substantial training data to overcome this issue (Schuller et al., 2011).

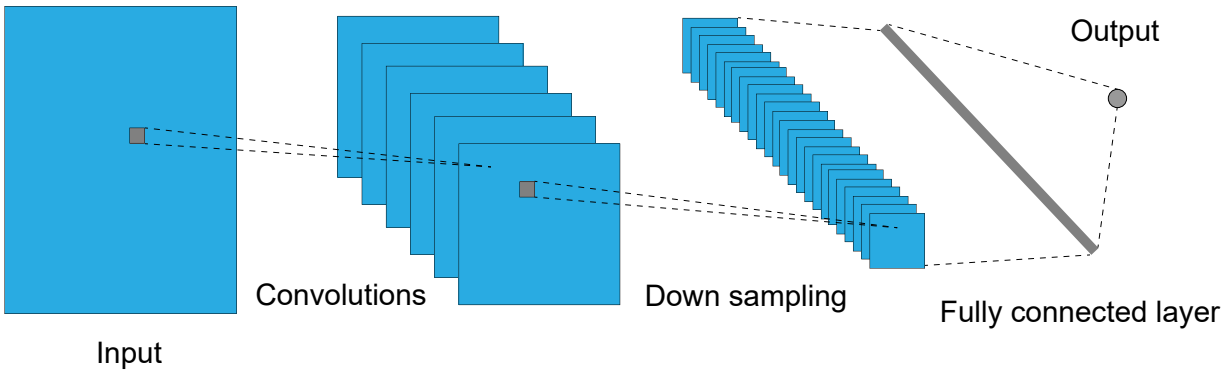

Figure 2.3: The figure shows the architecture of a convolutional neural network (CNN).

Convolutional Neural Networks

Convolutional Neural Networks (CNNs) represent a specialized type of ANN that mitigates overfitting issues, which commonly exist in the MLP. CNNs achieve this by simplifying the input through convolutional operations that extract smaller and more manageable features. This approach is inspired by the biological functioning within the animal visual cortex, where individual neurons are activated by stimuli within a limited region of the visual field. CNNs are widely used in various applications, such as medical imaging (Basu et al., 2017).

As represented in Figure 2.3, a CNN is composed of an input layer, multiple hidden layers, and an output layer. The hidden layers include a series of convolutional layers, pooling layers and a fully connected layer. The convolutional layers perform convolution operations on the input data with convolutional kernels or filters. The outcomes of these operations, referred to as the feature maps, undergo a non-linear activation function (typically ReLu) before feeding to the subsequent layer.

Following the convolutional layers, pooling layers serve to keep crucial features while reducing the input dimensions for the next convolutional layer. Average pooling and max pooling are two types of pooling strategies that are commonly used. After several iterations of convolution and pooling, a fully connected layer transforms the 2-dimensional feature map into a 1-dimensional feature vector, generating the output for classification or regression.

In the field of automatic depression detection, researchers typically utilize spectrograms derived from a short-time Fourier transform or other images related to speech signals as input data (Huang, Epps, Joachim, Stasak, et al., 2020; Saidi et al., 2020; Y. Wang et al., 2022; Q. Wang & Liu, 2022), which have demonstrated the effectiveness of CNNs. For example, Saidi et al. (2020) introduced a CNN-based methodology for automatic feature extraction in depression detection. In a similar vein, Y. Wang et al. (2022) extracted the deep features of MFCC feature maps by using CNN, achieving a depression detection accuracy of 77%. Subsequently, Q. Wang & Liu (2022) proposed an approach that uses more sophisticated deep spectral

features and modulates the architecture of CNN to achieve a better accuracy of 82.7% in depression detection.

However, the main limitation of CNNs used for speech data is that it does not take the sequential context of speech into account, thus likely to lose the information in the temporal respect.

Long Short-Term Memory

Long Short-Term Memory Networks (LSTMs), as a type of recurrent neural network (RNN), was first proposed by Hochreiter & Schmidhuber (1997). Although conventional RNNs can use their cyclic links to capture a certain amount of context and store temporal representations from previous inputs, they suffer from the vanishing gradient issue (Bengio et al., 1994). Therefore, LSTM is expected to solve this problem, providing a short-term memory for RNN that can persist over a large number of timesteps and can remember the optimal amount of contextual information benefits to the classification task, thus representing "long short-term memory". They find extensive use in time series data processing and prediction, with particular effectiveness in speech signal applications.

An LSTM network shares similarities with an RNN, with the key difference being the replacement of the nonlinear hidden units with memory units. A single LSTM unit consists of a cell, a forget gate, an input gate and an output gate. These three gates work together to update the cell state. The forget gate in an LSTM cell decides the portion of redundant information discarded from the previous timestep, such a gate has been demonstrated its effectiveness (Gers et al., 2000). The forget gate in the cell C at the time t is defined as follows:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f), \tag{2.12}$$

where W_f is the weight, b_f is the bias and f_t is the outcome of forget gate that is activated by a non-linear activation function σ of the weighted sum of the input vector x_t in the current cell and the state vector h_{t-1} in the previous cell.

The input gate decides the portion of new information learning and storage in the new memory cell. The input gate i_t in the cell C at time t is defined as follows:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i), \tag{2.13}$$

where W_i is the weight, b_i is the bias and the rest is identical to Equation 2.12. The candidate information that might be updated in the state are represented as follows:

$$\hat{c}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c), \tag{2.14}$$

where W_c is the weight and b_c is the bias. The $\hat{c}_t$ determines the portion of the input that can be

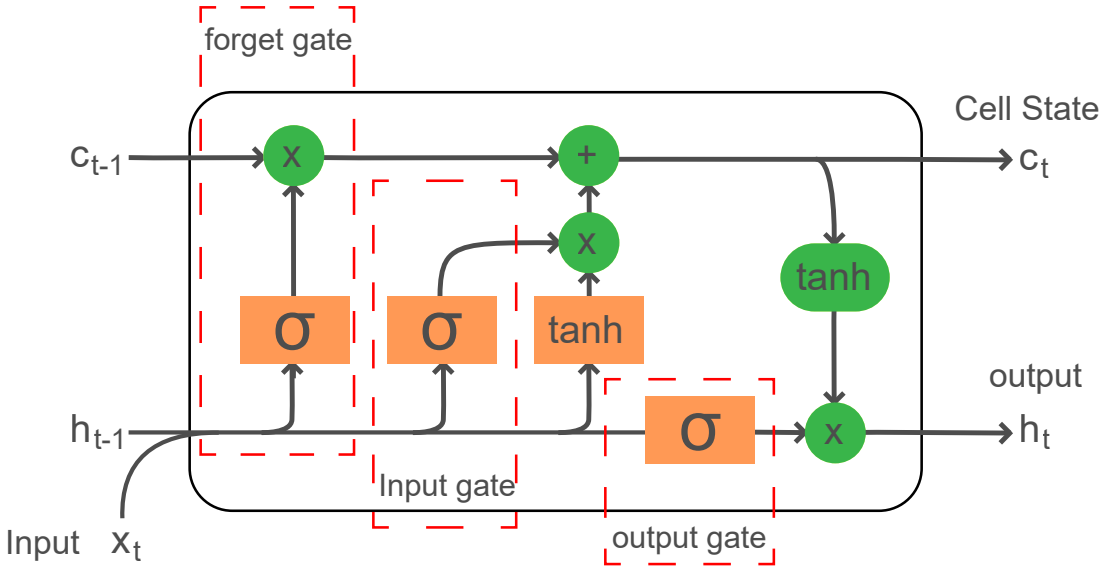

Figure 2.4: The figure presents the architecture of an LSTM cell at time t, including the forget gate, input gate, and output gate.

updated in the new cell state. The new cell state is then updated with the forget and input gates as follows:

$$c_t = f_t \odot h_{t-1} + i_t \odot \hat{c}_t, \tag{2.15}$$

where the operator $\odot$ is the Hadamard product, meaning the element-wise product. Here, c_t can be considered as output for a cell at time t. An output gate is defined to control the conversion of a cell output c_t to the hidden state h_t, as follows:

$$o_t = \sigma\left(W_o \cdot [h_{t-1}, x_t] + b_o\right). \tag{2.16}$$

The hidden state h_t at time t can then be defined as follows:

$$h_t = o_t \odot \tanh(c_t) \tag{2.17}$$

where tanh is a pointwise non-linear tangent activation function.

These gates are able to store and access information over lengthy sequences, by updating the cell state by taking the output from the previously hidden layer and the current input into account. This makes LSTMs particularly apt for processing speech signals. Consequently, in the domain of depression detection based on speech, LSTMs are a prevalent method (Z. Zhao et al., 2019, 2020; Y. Zhao, Liang, et al., 2021; Y. Zhao, Xie, et al., 2021; Dumpala et al., 2022; Muzammel et al., 2021; Rejaibi et al., 2022; Tao, Ge, et al., 2023). For example, Rejaibi et al. (2022) utilized LSTMs in conjunction with MFCC features to not only detect depressed individuals but also predict the severity of their condition. Moreover, Z. Zhao et al. (2020) introduced an additional attention layer to the LSTM architecture, enabling the transfer of knowledge learned

from the speech recognition task to the depression recognition task. The results were reported to outperform other speech-based systems with a Root Mean Square Error of 5.51 on the Patient Health Questionnaire-8 scale. Building upon this, Tao, Ge, et al. (2023) proposed a novel architecture combining LSTMs and multi-local attention mechanisms, achieving improved performance and efficiency in detecting depressed speakers.

2.3.5 Attention Mechanisms for Depression Detection

Recently, deep neural networks have become popular for the task of depression analysis (Aloshban et al., 2020, 2022; Wu et al., 2022; Du et al., 2023). Typically, the raw audio signal or extracted hand-crafted features of depression are used as input and fed to the deep neural networks for detection. For instance, researchers proposed feeding a combination of hand-crafted and deep-learned features, i.e., spectrogram information, to convolution neural networks (CNNs) to infer the severity of depression from speech (He & Cao, 2018). Meanwhile, Mel Frequency Cepstrum Coefficients are fed to Recurrent Neural Networks (RNN), a model dealing effectively with the sequential structure of speech, also showing a robust performance in depression detection (Rejaibi et al., 2022). However, these deep networks struggle when recordings are longer than a few seconds and give rise to thousands of feature vectors, possibly weakly labelled. The attention mechanisms (Vaswani et al., 2017), a group of algorithms in the context of deep learning, can be used to address this problem.

The use of attention mechanisms has become increasingly prevalent in the deep learning community, and they have become successful in a broad range of fields including speech recognition (Yeh et al., 2019), image retrieval (Ge et al., 2023), video description (Zhu & Jiang, 2019), and more recently, in automatic depression detection (Z. Zhao et al., 2019; Niu et al., 2020; Yin et al., 2023). The attention mechanisms can emphasize the most relevant information by drawing global dependencies between input and output. For instance, (C. Cai et al., 2021) proposed attention-based CNNs to focus on time-domain speech signals to obtain multiscale contextual information related to depression for classification. However, as mentioned earlier, these approaches lack attention to different types of speech thus neglecting the information specific to each of them. To the best of our knowledge, no previous work has employed attention mechanisms to extract complementary information from different types of speech to improve depression detection.

Attention mechanism led to the development of hierarchical attention networks that are capable of emphasizing information being modelled at different levels in the hierarchical structures by leveraging more than one level of attention in the networks (Z. Yang et al., 2016; Ge et al., 2021; Mallol-Ragolta et al., 2019). Based on this, researchers proposed hierarchical attention transfer networks that learn attention from a speech recognition task and then transfer attended linguistic information hierarchically to measure depression severity (Z. Zhao et al., 2019, 2020). However, such teacher-student frameworks require defining a specific task as the "teacher" to

teach the other task. This definition is typically based on empirical knowledge, leading to difficulty in finding the best task as the teacher. Compared to these approaches, we propose a student-student hierarchical framework that uses read and spontaneous speech of the same speaker to learn from each other to emphasize common information. To the best of our knowledge, no existing work has yet explored such a framework to enhance common information of two types of speech for depression detection.

2.3.6 Evaluation

The evaluation of a depression detection system is crucial to assess the quality of a designed detection approach. In this subsection, we will discuss how to assess the performance of different depression detection systems by utilizing appropriate evaluation approaches.

Evaluation Protocols

In the context of evaluation, the entire dataset is commonly split into disjointed subsets, namely *training sets* and *test sets*. The training sets are used to produce a trained model, while the test sets are used to evaluate the performance of the model. In general, there are two approaches to the process of data division:

- k-fold cross-validation: the entire dataset is randomly and evenly split into k disjoint subsets, $k-1$ subsets are used as training sets while the rest of one single subset is used as the test set for the model. Such a process is repeated for k times, a left-out subset as the test set at each time. As a result, the model will be trained for k times, and the averaged score on the test set can be considered as the final result.

- Leave-one-out cross-validation: is a particular type of k-fold cross-validation. This approach is also often used in depression detection when the amount of data is limited due to the difficulty in recruiting depressed patients. In that case, only one sample is used for test while the remaining data is used for training. This process is repeated for N times where N is the number of samples. At each time, an unused sample is selected for test. The averaged score is also the final result.

Numerous research studies primarily present the best singular performance outcome, which fails to represent the distribution of the performance of the model. This is mainly because the weights in the hidden layer are random in the initialization of the model training. This may lead to potential suboptimal results across training instances, although the same parameters are utilized in the model. Therefore, the training process is typically repeated several times using the same experimental design to obtain the overall model performance. In this thesis, we reported the average and standard deviation of the performance in all experiments over at least 10 times repetitions.

Evaluation Metrics

The evaluations of depression detection are usually defined as the difference between the predictions and the ground truth, and such evaluations varied across tasks. For the task of assessing the severity of depressed individuals, The evaluation metrics include but are not limited to root mean squared error (RMSE) and mean absolute error (MAE) (Chai & Draxler, 2014). In this thesis, we do not focus on the assessment of the severity of depressed speakers, and thereby we do not use these evaluations.

While for the task of recognizing depressed speakers, the commonly used evaluation metrics include Accuracy, Precision, Recall and F1-score, defined as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{2.18}$$

$$Precision = \frac{TP}{TP+FP} \tag{2.19}$$

$$Recall = \frac{TP}{TP+FN} \tag{2.20}$$

$$F1-score = 2*\frac{Precision*Recall}{Precision+Recall} \tag{2.21}$$

where TP is the true positive, FP is the false positive, TN is the true negative and FN is the false negative. In this thesis, these parameters are defined as follows:

- TP: number of depressed patients that are correctly classified as depressed;
- FP: Type I error, meaning the number of non-depressed individuals that are misclassified as depressed;
- TN: number of non-depressed individuals that are classified correctly as non-depressed;
- FN: Type II error, meaning the number of depressed individuals that are misclassified as non-depressed;

Typically, Accuracy is the most widely used evaluation in the field of machine learning. However, the prevalent problem in many public depression speech datasets is that the number of recordings for depressed speakers is significantly less than those for non-depressed speakers. This imbalanced class distribution may lead to an unreliable accuracy in which a classifier always assigns unseen samples to the majority class, leading to high accuracy. Therefore, numerous studies also take Precision, Recall and F1-score into account for the evaluations. These evaluation metrics not only consider accuracy, but also take the situations of Type I error and Type II error into account. In the field of depression detection, Type II error usually lead to more serious misclassification costs, because this could imply overlooking a diagnosis of depression, leading to delayed treatment or, in worse scenarios, potential harm to patients, including the risk of suicide. On the contrary, a Type I error simply means providing additional medical attention

to someone who is not actually in a critical condition. Therefore, Recall is more practical than Precision in clinical depression diagnosis. Most studies tend to report F1-score that takes both Recall and Precision into account.

2.3.7 Conclusions

In this chapter, we provided an overview of the background in the context of automatic depression detection. Firstly, We introduced the cognitive alterations observed in the depressed patients, exploring how these changes manifest in speech production from both acoustic and linguistic perspectives. Subsequently, we introduced widely employed acoustic and linguistic features for detecting depression. We reviewed the studies in investigating depressed speech to further confirm such speech features are effective and can be used to distinguish between depressed and non-depressed speakers.

Next, we present a review of the automatic depression detection pipeline, including data preprocessing and subsequent recognition stages. Then we reviewed the algorithms that are commonly utilized in the automatic depression detection studies, including Support Vector Machines, Random Forests, Multi-layer Perceptrons, Convolutional Neural Networks, and Long Short-Term Memory Networks. Furthermore, we introduced evaluation methods extensively applied to assess the effectiveness of automatic depression detection systems. Finally, we highlight the main advantages in the use of such systems.

In the next chapter, we will introduce the dataset collected for this thesis and show the advantages of our dataset by comparing to other publicly available datasets. Additionally, we establish baseline models using our dataset, enabling meaningful comparisons with the experimental results detailed in subsequent chapters.

Chapter 3

The Androids Corpus: A New Publicly Available Benchmark

In this chapter, we introduce the dataset that was used in the experiments presented in the rest of the thesis. A novel benchmark, the Androids Corpus, for automatic detection of depression using speech. The corpus includes 118 participants, out of which 64 were diagnosed with depression by professional psychiatrists. Each participant has at least one recording, resulting in a total of 228 recordings.

The Androids Corpus (https://github.com/androidscorpus/data) consists of two types of speech samples: 112 recordings of read speech, where each speaker reads the same text, and 116 recordings of spontaneous speech, where all speakers answer the same questions in the same order presented by an interviewer. Both types of speech for 110 speakers, including 58 diagnosed with depression, are available in the corpus, while for the remaining 8 speakers, only either read or spontaneous speech is available.

In terms of the corpus duration, the total recording time for read and spontaneous speech is 1 hour, 33 minutes, and 49 seconds, and 7 hours, 24 minutes, and 22 seconds, respectively. Along with the data, we also provide experimental protocols and baselines to ensure reproducibility of the results.

3.1 Introduction

In order to address the issues discussed in Chapter 1, speech-based depression detection has become increasingly popular, and the computing community has made significant efforts to develop methods for detecting depression, as seen in studies such as (L. Yang et al., 2021; Cummins et al., 2020; Rejaibi et al., 2022; Alghowinem, Goecke, Wagner, Parkerx, & Breakspear, 2013; Y. Yang et al., 2012; Alsarrani et al., 2022). In particular, this topic has been presented several times at the Audio-Visual Emotion Challenge (AVEC), a competitive event established with the objective of evaluating multimedia processing and machine learning techniques for au-

tomated analysis of emotions in audio, visual, and audio-visual media (Valstar et al., 2013, 2014, 2016; Ringeval et al., 2017, 2019). Encouragingly, some studies have reported achieving depression detection results comparable to those of General Practitioners (Aloshban et al., 2022), who are typically the initial point of contact for depression intervention in many healthcare systems. However, progress in this field continues to be impeded by a lack of data. One reason for this is that sophisticated techniques, particularly those leveraging on deep neural networks, require increasingly large amounts of data for effective training. Another reason is that, without multiple corpora, it is impossible to test the robustness of a given approach to various changes in factors such as settings, sensors, people, language, and other corpus-specific characteristics.

For these reasons, researchers have constructed various datasets (see Table 3.1), typically collecting audio files when participants perform read and free/spontaneous speech. Read and spontaneous speech are two different types of data that represent planned and unplanned speech, respectively. Interviews are widely used for collecting spontaneous speech, although some studies also use picture description tasks (H. Cai et al., 2020; Tasnim et al., 2022), both of these tasks can be used to obtain spontaneous speech.

The dataset in Mundt et al. (2007) was constructed by involving 35 patients who were referred by their treating physicians and were beginning psychotherapy treatment for depression. The participants were instructed to call a telephone number once a week for six weeks, and their speech data was automatically collected via the telephone using an Interactive Voice Response (IVR) system. In addition, the participants were assessed weekly for six weeks using standard depression severity measurements, Hamilton Rating Scale for Depression (HRSD) and Quick Inventory of Depressive Symptomology (QIDS). The speech samples include automatic speech production tasks, such as counting, reciting the alphabet, and passage reading, as well as free speech tasks. The study examines the relationships between voice acoustic measures and depression severity, while also evaluating any changes in voice measures that may be associated with treatment response. Several subsequent early studies of speech-based depression detection are based on this dataset (Sturim et al., 2011; Trevino et al., 2011; Quatieri & Malyska, 2012; Cummins, Epps, & Ambikairajah, 2013; Cummins, Epps, Sethu, et al., 2013; Helfer et al., 2013).

The Datasets of AVEC 2013 (Valstar et al., 2013) and AVEC 2014 (Valstar et al., 2014) are subsets of the Audio-Visual Depression Language Corpus (AViD-Corpus), which consists of 292 participants and a total of 340 video clips recorded via a webcam and a microphone during a Human-Computer Interaction task. Each participant performed free speech, reading, singing, and picture description tasks, while their self-reported depression severity level, as measured by the Beck Depression Inventory-II (BDI-II), was used as groundtruth. The AVEC 2014 dataset (Valstar et al., 2014) is a subset of AVEC 2013 and comprises 300 videos obtained from the previous Human-Computer Interaction task in AVEC 2013. However, AVEC 2014 only includes two out of the original 14 tasks, namely read speech (reading aloud a passage, such as

Table 3.1: The survey of representative depression speech datasets. Abbreviations: PT - Depressed patients, HC - Healthy controls, HRSD – Hamilton Rating Scale for Depression, QIDS – Quick Inventory of Depressive Symptomology, BDI – Beck Depression Inventory, PHQ-9 – Patient Health Questionnaire, DSM - Diagnostic and Statistical Manual of Mental Disorders, SDS - Zung Self-Rating Depression Scale, GAD-7 - Generalized Anxiety Disorder scale.

Datasets	Types		Subjects	Labels	Duration	Language
	Read	Spontaneous				
Mundt et al. (2007)	√	√	35 PT	HRSD QIDS	-	English
Valstar et al. (2013)	√	√	84 PT 208 HC	BDI-II	240 hours	German
Valstar et al. (2014)	√	√	84 PT 66 HC	BDI-II	240 hours	German
Gratch et al. (2014)		√	37 PT 152 HC	PHQ-8	-	English
DeVault et al. (2014)		√	275 participants	PHQ-8	73 hours	English
H. Cai et al. (2020)	√	√	23 PT 29 HC	HRSD DSM-IV	-	Chinese
Y. Shen et al. (2022)		√	30 PT 132 HC	SDS	2.26 hours	Chinese
Tasnim et al. (2022)		√	312 PT 240 HC	PHQ-9 GAD-7	-	English

"The Wind of the North and the Sun") and free speech (answering one of several questions). This selection of tasks allows for a more focused investigation of affect and depression analysis.

The Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ) (Gratch et al., 2014) is a collection of clinical interviews that aim at identifying psychological distress conditions including anxiety, depression, and post-traumatic stress disorder (PTSD). The corpus includes audio and video recordings of 189 participants and each participant was interviewed with an animated virtual interviewer to answer a standardized set of questions, while PHQ-8 was used

to annotate depression patients. This dataset was explored in the AVEC 2016 (Valstar et al., 2016) and AVEC 2017 (Ringeval et al., 2017). E-DAIC (DeVault et al., 2014) is an extended dataset that builds on the DAIC-WOZ and encompasses 275 participants. The data was obtained through semi-clinical interviews intended to aid in diagnosing psychological distress conditions. The dataset was annotated with PHQ-8 to be in line with the evaluation of DAIC-WOZ, and it was explored in the AVEC 2019 (Ringeval et al., 2019).

The multimodal open dataset for mental-disorder analysis (MODMA) (H. Cai et al., 2020) includes audio recordings, EEG data and questionnaires from 52 participants, of whom 23 participants were diagnosed with major depressive disorder (MDD) by at least one clinical psychiatrist. Each participant performed an interview, word reading, passage reading and picture description task. The audio recordings of the interview consist of answers to 18 questions based on the DSM-IV and HRSD scales.

The Emotional Audio-Textual Depression Corpus (EATD-Corpus) (Y. Shen et al., 2022) allows depression detection based on speech characteristics and linguistic content. The dataset consists of audio and text transcripts of interviews with 162 Chinese students, of whom 30 participants were regarded as depressed. The identification of depressed speakers was based on their score on the SDS scale, which indicates the severity of depression. In the interview, each participant was asked to answer three randomly selected questions by a virtual interviewer in an APP. Their audio responses were then collected and uploaded online.

The Depression and Anxiety Crowdsourced Corpus (DEPAC) (Tasnim et al., 2022) aims to provide a large collection of audio recordings of individuals with varying degrees of depression and anxiety. The dataset consists of 2674 audio clips of 571 participants, and each paricipant was asked to perform 5 tasks, i.e. phoneme pronunciation, phonemic fluency test, picture description, semantic fluency test and prompted narrative task. Each participant was annotated with both PHQ-9 and GAD-7 to indicate the severity of depression and anxiety, out of which 312 participants were labeled as depression patients (total 552 participants had results of the questionnaire indicating depression). Additionally, 139 of the same 571 participants were annotated with anxiety.

While several datasets have been provided to researchers for automatic depression detection, it is evident that these corpora suffer from certain limitations. One such limitation for the dataset (Mundt et al., 2007) is the lack of healthy control data for comparison, which makes it difficult to identify whether a speaker is depressed or not. Moreover, the datasets for AVEC 2013, AVEC 2014, DAIC-WOZ, E-DAIC, and EATD are unbalanced and some of them only include one type of speech. Another issue with the annotation of participants in these datasets, as well as in DEPAC, is that it is based on self-reported questionnaires. This approach may introduce a subjective bias, due to misalignment between clinicians' observations and patients' responses (R. P. Greenberg et al., 1992). Additionally, the number of participants in the MODMA dataset is limited, which hinders the training of deep learning approaches and limits the model's

ability to generalize.

For the reasons above, this chapter introduces the Androids Corpus, which serves as a novel benchmark for speech-based depression detection and is publicly available. The corpus was specifically designed to offer distinct and complementary opportunities in comparison to other available corpora.

- The labels for depressed and non-depressed speakers were given by professional psychiatrists and not through the rates of self-assessment questionnaires. This means that the data allows one to investigate depression detection and not prediction of scores obtained through questionnaires, known to be affected by multiple biases (Paulhus & Vazire, 2007).

- The data were collected with the microphone of a laptop in the mental health centers where the depression patients are treated. Such an *in-the-wild* setting corresponds to the situation in which doctors and depressed participants meet for their therapeutic interactions. This allows one to develop an approach in conditions representative of a real-world application environment.

- The corpus includes both *read* and *spontaneous* speech recordings for 110 out of a total of 118 participants. This enables investigation of differences in speech properties between different types of speech from the same speaker, and exploration of the possibility of detecting depression by using multiple tasks.

- The populations of depressed and non-depressed participants (64 and 54 individuals, respectively) have the same distribution in terms of age, gender and education level. This limits speaking differences resulting from factors other than depression.

- In the case of spontaneous speech (collected during interviews), the corpus includes a manual segmentation into turns. This allows one to investigate conversational dynamics and turn-taking related features (e.g., turn length distribution, speaking time, etc.). In addition, it excludes the voice of the interviewer from the conversation, thus eliminating the impact of the voice of the interviewer on depression detection when using spontaneous speech.

- The corpus includes reproducible experimental protocols (the split of the participants into subsets to be used for a k-fold setup). This allows rigorous comparisons across multiple results in this thesis obtained over the data.

- The corpus provides an OpenSmile (Eyben et al., 2013) configuration file allowing one to extract a standard feature set from the speech recordings. This improves the reproducibility of the results obtained over the data.

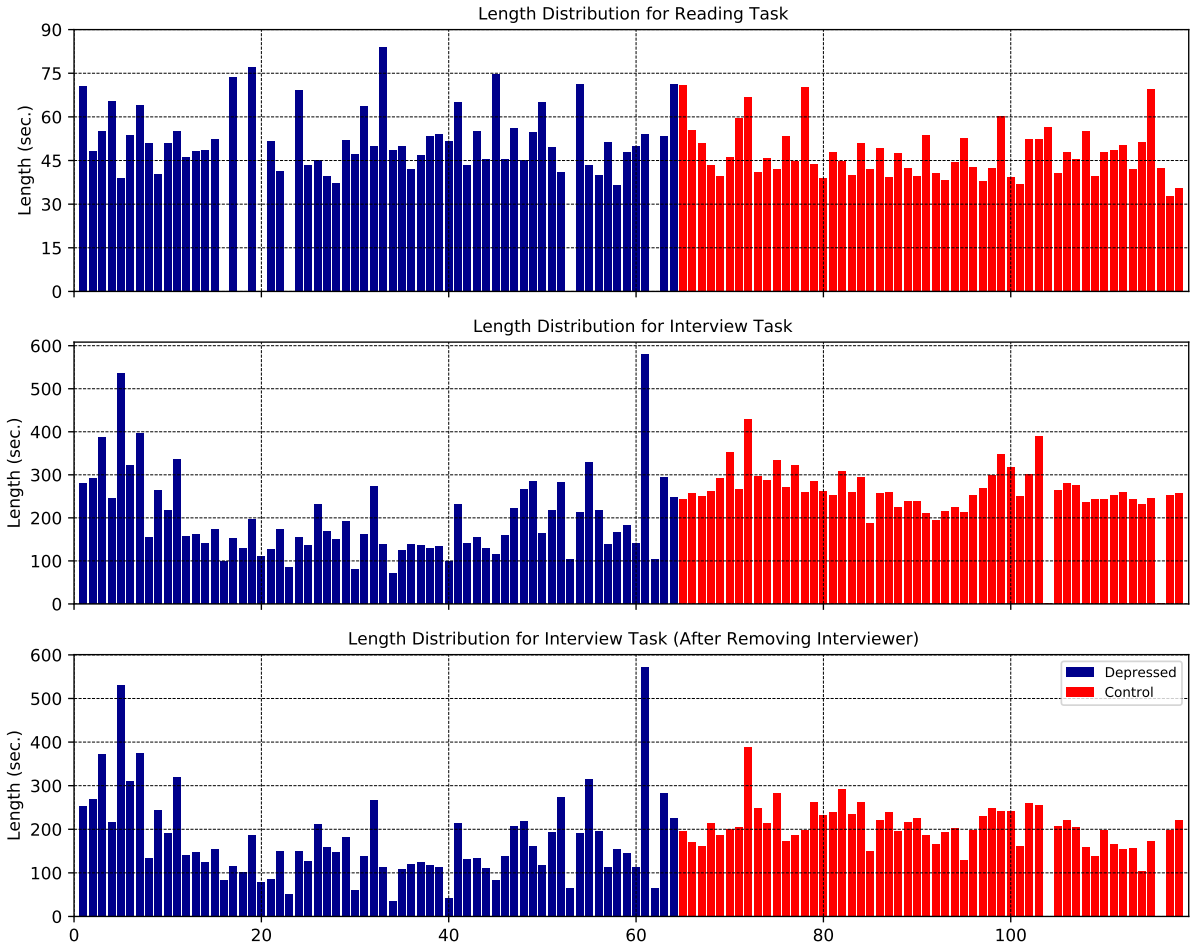

Figure 3.1: Length distribution across participants. The top bar chart shows the length of the RT recordings for each participant, the middle one shows the same information for the IT recordings and the bottom one does the same for the IT data after removing the turns of the interviewer. Missing bars correspond to participants that did not perform one of the tasks.

- The data is in Italian, a language not widely represented in other publicly available corpora. This allows one to test cultural effects and to investigate whether approaches developed in one language can extend to other languages.

Given the diversity and depth of the corpus, there is a possibility of uncovering new research directions that were not previously anticipated. Therefore, the list above should not be considered exhaustive.

The rest of this chapter is organized as follows: Section 3.2 describes the corpus in detail, Section 3.3 presents the baseline results and the final Section 3.4 draws some conclusions.

3.2 The Androids Corpus

The corpus consists of recordings from 118 participants, including 64 participants diagnosed with depression (referred to as *depression participants* hereafter) and 54 participants without any history of mental health issues (referred to as *control participants* hereafter). Each participant was asked to complete two tasks: the *Reading Task* (RT) and the *Interview Task* (IT). The Reading Task required participants to read a short fairy tale called "*The Wind of the North and the Sun*" by Aesop. Such a text was extracted from a book for children because it is plain and easy to read. This ensures that the participants can perform the Reading task independently of their education level and condition (depressed or non-depressed). The IT involved answering questions about daily life, such as "*What did you do last week end?*" posed by an interviewer who was instructed to speak as little as possible. Such a task replicates a typical therapeutic interview with the psychiatrists of five Mental Health Centres involved in the study. The recordings of the IT were manually segmented to discard the voice of the interviewer. The recordings of both RT and IT were collected in standard clinical consultation rooms of the Mental Health Centers above. The data were recorded with a standard laptop microphone (participants and interviewers were always at the same distance from the microphone). The aim of the RT and IT was to produce recordings of read and spontaneous speech, respectively.

The data collection involved psychiatrists to diagnose the patients with the *Diagnostic and Statistical Manual of Mental Disorders 5* (DSM-5). The number of patients with specific pathologies are as follows: 22 cases of major depressive disorder, 15 cases of bipolar disorder in the depressive phase or with the last depressive episode, 8 cases of reactive depression, 7 cases of endo-reactive depression, 5 cases of anxiety-depressive disorder, and 1 case of persistent depressive disorder. For the remaining 6 depression participants, no specific pathology was identified. The distribution of diagnoses indicates that there is sufficient diversity to cover all aspects of depression and not just certain types.

All participants were involved on a voluntary basis and they all signed an informed consent letter formulated in accord with Italian and European privacy and data protection laws[1]. The data collection followed the ethical regulations of countries and institutions involved in the work. In the case of the depression group, the participants were recorded while undergoing treatment, which means that they were experiencing the pathology. This makes it possible to detect depression traces among their speech samples. In the case of the control group, the participants were recruited through word of mouth and selected based on their age, gender, and education level to match the distribution of the depression group (see below).

Table 3.3 presents demographic information about the participants. Most of them (110 out of 118) participated in both RT and IT, whereas a few only participated in one of the two (see Table 3.2). This is the reason why Table 3.3 provides information not only for the entire set of

[1]The ethical committee of the Department of Psychology at Universitá degli Studi della Campania, "Luigi Vanvitelli", authorized the experiment with protocol number 09/2016.

Table 3.2: Participant distribution across tasks. Acronyms RT and IT stand for *Reading Task* and *Interview Task*, respectvely.

Group	Only RT	Only IT	RT and IT	Total
Control	2	0	52	54
Depression	0	6	58	64
Total	2	6	110	118

Table 3.3: Demographic information. Acronyms *F* and *M* stand for Female and Male, respectively. Acronyms *L* and *H* stand for Low (8 years of study at most) and High (at least 13 years of study) education level, respectively. The sum over the education level columns does not correspond to the total number of participants (118) because 2 of these did not provide details about their studies.

Task		Age	M	F	L	H
RT	Control	47.1 ± 12.8	12	42	19	35
	Depression	47.4 ± 11.9	20	38	25	32
	Total	47.2 ± 12.3	32	80	44	67
IT	Control	47.3 ± 12.7	11	41	19	33
	Depression	47.5 ± 11.6	21	43	29	33
	Total	47.4 ± 12.1	32	84	48	66
Total	Control	47.3 ± 12.7	11	41	19	33
	Depression	47.4 ± 11.9	20	38	25	32
	Overall	47.3 ± 12.2	31	79	44	65

participants, but also separately for the two tasks. There is no significant difference observed between depression and control participants in terms of gender and education level distribution according to a χ^2 test ($p > 0.05$). This applies to the entire corpus as well as to the subsets of participants involved in each task. Similarly, according to a two-tailed t-test ($p > 0.05$), there is no statistically significant difference in age between the depression and the control group. Overall, this ensures that the differences in speech between the two groups are due to the pathology rather than other factors that might have an impact on speech. The female participants are roughly 2.5 times more than the male participants, which is consistent with epidemiological observations that females are more likely to experience depression than males roughly in the same proportion (Andrade et al., 2003; R. Kessler et al., 2003).

Figure 3.1 shows the length of the recordings for each individual. The total duration of the RT recordings is 1 hour, 33 minutes and 49 seconds, while the average duration is 50.3 ± 10.3 seconds. When considering separately depression and control participants, the averages are 52.9 $\pm$ 10.9 and 47.4 ± 8.8 seconds, respectively. The difference is statistically significant ($p < 0.01$ according to a two-tailed t-test), in line with previous studies showing that depression patients tend to read slower (Tao et al., 2020). In the case of the IT recordings, the total duration is 7

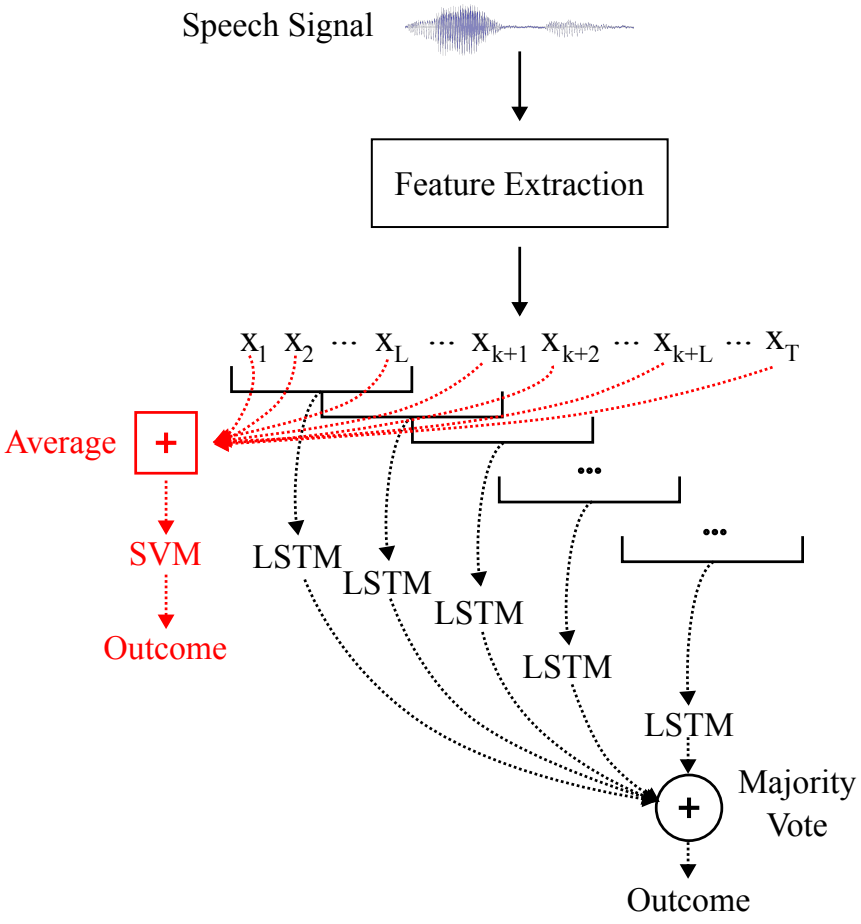

Figure 3.2: Baseline approaches. The diagrams shows the two baseline approaches used in the experiments. The symbol $\boxplus$ corresponds to the average, while the symbol $\oplus$ corresponds to the majority vote.

hours, 24 minutes and 22 seconds (the overall average and standard deviation are 229.8 $\pm$ 86.6 seconds). According to a two-tailed t-test, there is no statistically significant difference between depression and control participants (the averages are 198.8 $\pm$ 99.2 and 268.0 $\pm$ 45.3 seconds, respectively). In the IT recordings, the total duration of the turns for the participants is 6 hours, 8 minutes and 21.8 seconds, for an average of 190.5 $\pm$ 84.1 seconds. The average lengths for depression and control participants are 176.4 $\pm$ 103.1 and 207.9 $\pm$ 47.3 seconds, respectively. Such a difference is statistically significant according to a two-tailed t-test ($p < 0.05$).

3.3 Baseline Approaches and Results

The goal of this section is to provide an initial set of results that can be used as a term of comparison to the experiments in the following chapters based on the corpus data. The baseline approaches (see Figure 3.2) include three main steps, namely *feature extraction*, *depression detection* and *aggregation*.

3.3.1 Feature Extraction

The initial step is the same for both baseline approaches. The goal of this step is to convert the original speech signals into sequences of feature vectors, a more suitable format for classification. In the baseline approaches, the extraction was performed with OpenSMILE (Eyben et al., 2013), an open-source tool widely used for the inference of social and psychological information in speech. The vectors were extracted from 25 ms long analysis windows at regular time steps of 10 ms. The window length is in line with the indications of the literature showing that values between 20 and 40 ms lead to satisfactory results in the inference of emotional and social phenomena (Schuller & Batliner, 2013; Schuller et al., 2009; Pappagari et al., 2020; Tayarani et al., 2019). These values are standard in the literature and were set a-priori. No attempt was made to find values that might lead to better performance. Every analysis window is converted into a feature vector $\vec{x}$ with OpenSmile (Eyben et al., 2013).

The features include *Root Mean Square of Energy* (RMSE), *Mel Frequency Cepstrum Coefficients* (MFCC) 1 to 12, *Zero Crossing Rate* (ZCR), *Voicing Probability* (VP) and *fundamental frequency* or *pitch* (F0);

- *Root Mean Square of the Energy* (1 feature): it provides a measure of how loud someone speaks and it was shown to account for depression in the literature (L.-S. A. Low et al., 2010; Schuller & Batliner, 2013).

  The definition of the physical energy E of the speech signal $s(k)$ is as follows (Kießling, 1997):

  $$E = \sum_{k=-\infty}^{+\infty} s^2(k) \tag{3.1}$$

  In particular, the energy $E(t)$ of a frame at time t is defined as follows:

  $$E(t) = \sum_{k=t-\frac{N}{2}}^{t+\frac{N}{2}+1} [s(k)w(t-k-1)]^2 \tag{3.2}$$

  where $w(k)$ is a window function for non-zero points from $t-N/2$ to $t+N/2+1$. This allows to define the root means square (RMS) energy of $E(t)$, which is used in this thesis for feature extraction.

- *Mel-Frequency Cepstral Coefficients 1-12* (12 features): they model the human peripheral auditory system that separates sounds according to their frequency components and transforms sound signals into neural signals for brain processing. Given that the human auditory system is more sensitive to sounds in low frequency than to sounds in high frequency, the Mel-Frequency Cepstral Coefficients (MFCCs) model these characteristics using a nonlinear scale, known as the Mel scale. The transform from a linear frequency

scale f to Mel scale $Mel(f)$ is defined as follows (Schuller & Batliner, 2013):

$$Mel(f) = 2595 \cdot \log\left(1 + \frac{f}{700}\right) \tag{3.3}$$

The most commonly used coefficients are from 0 to 16. In particular, the coefficients 0-12 are the most frequently used in speech recognition. Given coefficient 0 describes the signal energy, we discard coefficient 0 from the feature set in order to avoid redundancy with the RMS of the Energy. Coefficients 1–12 provide a measure of the phonetic content of the speech signal and they were shown to be effective in a wide spectrum of applications aimed at inferring social and psychological information from speech, including depression (Cummins, Scherer, et al., 2015);

- *Fundamental Frequency* (1 feature): the fundamental frequency $F0$ plays an important role among speech parameters. This is mainly because, compared to other speech parameters, the changes in the fundamental frequency are much easier perceived by humans (Schuller & Batliner, 2013). In addition, $F0$ is the frequency that carries the highest energy in the signal, known to convey depression-relevant information (France et al., 2000; Ozdas et al., 2004; Quatieri & Malyska, 2012). In this thesis, the $F0$ is extracted with the sub-harmonic summation approach in line with the study (Schuller & Batliner, 2013);

- *Zero-Crossing Rate* (1 feature): The zero crossing rate (ZCR) represents the frequency of zero crossings within a frame. In other words, it occurs when successive samples exhibit alternating algebraic signs in the context of discrete-time signals. ZCR is defined as follows (Schuller & Batliner, 2013):

$$ZCR(t) = \sum_{k=t-\frac{N}{2}}^{t+\frac{N}{2}+1} s_0(k), \tag{3.4}$$

$$s_0(k) = \begin{cases} 0 & \text{if } \operatorname{sgn}[s(k)] = \operatorname{sgn}[s(k-1)] \\ 1 & \text{if } \operatorname{sgn}[s(k)] \neq \operatorname{sgn}[s(k-1)] \end{cases} \tag{3.5}$$

where $s_0(k)$ is the signal contains points from $t - N/2$ to $t + N/2 + 1$. In the context of speech signal analysis, the ZCR serves as an indicator of the number of instances when the amplitude traverses a zero value within a prescribed temporal segment. By analyzing the proportion of zero crossings in a specific time interval, ZCR can provide information about the frequency distribution. In particular, it mainly correlates with the low-frequency components and it would be low when the signal contains strong low-frequency components (Schuller & Batliner, 2013). In addition, it is an alternative estimate of the funda-

mental frequency and it was shown to lead to 80% accuracy in predicting whether a naive listener considers a speaker depressed (Nilsonne & Sundberg, 1985);

- *Voicing probability* (1 feature): According to the study (Schuller & Batliner, 2013), voicing probability is obtained by using the autocorrelation function to calculate the first distinct peak in relation to the total signal energy. This feature accounts for the probability of an analysis window corresponding to the emission of voice and it was shown to be effective for the inference of affective and mental states (Cummins, Scherer, et al., 2015; Gobl & Chasaide, 2003).

The 16 features are enriched by taking into account the differences between the feature values extracted from two consecutive analysis windows, thus reaching a dimensionality $D = 32$. The feature set was initially designed for the Interspeech Emotion Recognition Challenge in 2009 (Schuller et al., 2009) and it has been widely used in various applications, particularly in the inference of social and psychological information from speech.

3.3.2 Depression Detection

This step allows to implement depression recognition based on extracted feature vectors. At the end of the feature extraction step, every recording is represented with a sequence of vectors $X = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_T\}$, where T is the total number of vectors (proportional to the length of the recording). The first baseline, referred to as BL_{SVM} hereafter, involves the calculation of the average of the $\vec{x}_k$ vectors and feeding it into a Support Vector Machine (SVM) with the linear kernel (see red components in Figure 3.2).

The second baseline (see Figure 3.2), referred to as BL_{LSTM} hereafter, involves the segmentation in which X is segmented into fixed-length subsequences of length $M = 128$, referred to as *frames*. The frames start at regular steps of length $M/2$, resulting in a half overlap between consecutive frames I_k, where k ranged from 1 to N (N is the total number of frames). Each frame is then used as input individually for a Long Short-Term Memory Network (LSTM) model and classified as either belonging to the depression or control group. The number of hidden states of one layer in LSTM is 32, and both parameters were set a-priori.

3.3.3 Aggregation

Given that there are multiple frames per recording for each speaker, there are multiple classification outcomes too for BL_{LSTM}. This makes necessary an *aggregation* step that takes as input the N classification outcomes (N is the total number of frames in a recording) and performs a majority vote, i.e., it assigns a recording to the class its frames are most frequently assigned to (see Figure 3.2):

$$\hat{c} = \arg\max_{c \in \mathscr{C}} n(c), \tag{3.6}$$

Table 3.4: Depression detection results. The table shows the results obtained over RT.

	Acc.	Prec.	Rec.	F1
Rand.	50.1	51.8	51.8	51.8
BL_{SVM}	69.7±6.6	69.6±3.5	76.3±12.9	71.9±4.8
BL_{LSTM}	83.4±2.6	85.0±3.5	83.0±2.3	83.8±2.1

Table 3.5: Depression detection results. The table shows the results obtained over IT.

	Acc.	Prec.	Rec.	F1
Rand.	50.5	55.2	55.2	55.2
BL_{SVM}	64.7±6.3	68.4±7.5	66.3±10.9	66.6±7.1
BL_{LSTM}	81.6±1.6	83.3±2.7	85.3±1.4	83.4±1.1

where $\mathscr{C}$ is the set of all possible classes (depression and control in the experiments of this work), $n(c)$ is the number of frames assigned to class c and $\hat{c}$ is the class assigned to the recording.

To provide a point of reference for comparison, all baselines are compared to a random classifier that assigns each recording to a specific class c based on the prior probability $p(c)$ of that class. The accuracy $\hat{\alpha}$ of such a classifier is defined as follows:

$$\hat{\alpha} = \sum_{c \in \mathscr{C}} p(c)^2 \tag{3.7}$$

where $\mathscr{C}$ is the set of all classes. Precision, recall, and F1 score of the random classifier are equivalent to the prior of the positive class, which in this thesis is the percentage of the number of depression participants.

3.3.4 Experiments and Results

All the baseline experiments were performed according to a k-fold protocol ($k = 3$) and the experiments were repeated k times. At every repetition, a different fold was used for the test while the remaining folds were used for training. All data from a given speaker is placed within the same fold. In other words, the experiments are designed to be *person independent*, where the data of the same individual does not appear in both the training and test sets. In this way, it ensures that the approaches detect depression and do not simply identify the speakers.

Our model is trained on a single Tesla T4 GPU with 16 GB memory. The Long Short-Term Memory networks (LSTMs) used in the study were configured with a number of hidden states of 32. The learning rate was set to 10^{-3}, and the number of training epochs was established at 100. The LSTMs were trained using RMSProp as an optimizer (Tieleman et al., 2012) with cross entropy (Rubinstein & Kroese, 2004) as a loss function. The experiments were repeated

$R = 10$ times due to the different random initialization of the weights of LSTMs in each fold. For this reason, the performance metrics are reported in terms of average and standard deviation across the R repetitions and k folds. The BS1 were implemented with scikit-learn (version in 0.23.2) (Pedregosa et al., 2011), while the BL_{LSTM} were implemented with PyTorch (version in 1.13.1+cu116) (Paszke et al., 2019).

Table 3.4 and Table 3.5 present performance metrics of depression detection including Accuracy, Precision, Recall and F1 Score obtained over the Read and Interview Tasks. All approaches show a statistically significant improvement over the random baseline ($p < 0.01$ according to a two-tailed t-test). These results may confirm that speech can be used for depression detection, and the baseline results might be comparable with *General Practitioners* (the accuracy ranges between 57.9% and 73.1%) (Mitchell et al., 2009). These baseline results can serve as a term of comparison for other approaches implemented in other chapters using the same data as well as the same experimental design.

3.4 Conclusions

In this chapter, we introduced the data used in the thesis. The data (Androids Corpus) is a new publicly available benchmark for speech-based depression detection. In addition to the description of the data, we surveyed the currently publicly available audio datasets used for depression detection, and we highlighted the advantages of the Androids Corpus by comparing it to those datasets.

Furthermore, we introduced speech features and experiment design of baselines used in the experiments in the later chapters to keep rigorous comparisons. Next, we provided baseline results with the predefined experimental protocols. In addition to depression detection, the Androids Corpus can be used to address various research problems such as identification of speech markers in read and spontaneous speech, analysis of interplay between read and spontaneous speech as well as the relationship between depression and conversational dynamics, etc. Therefore, this corpus offers researchers an opportunity to expand their research on automatic depression detection.

Chapter 4

Speech Duration and Silences for Depression Detection

4.1 Introduction

In the previous chapters, we reviewed the studies investigating automatic depression detection. The reason for focusing on such a technology is due to the alarming escalation in the number of depressed patients and the consequent economic burden on society in recent years. By serving as a supplementary tool for clinical diagnosis, automatic depression detection holds the potential to alleviate the situation to some extent. One important aspect of implementation, at least to a partial extent, is to identify depression markers, i.e., reliable, measurable and, possibly, machine-detectable traces of the pathology. In fact, the availability of such markers can make it easier for non-specialised doctors to effectively identify people affected by depression and, thereby leading to an increase in the percentage of cases that obtain psychiatric attention and proper treatment. Additionally, the traditional clinical diagnosis of depression primarily relies on questionnaires, such as *Hamilton Depression Rating Scale*. This way of diagnosis requires potential patients to make self-assessments towards their internal states. However, this seems to introduce various subjective biases, particularly when evaluating internal states (Shrout et al., 2018). These biases can arise from decreased cognitive ability of patients or the stigma associated with mental disorders. Furthermore, the assessments provided by clinicians are not immune to subjective biases either (Walfish et al., 2012), and therefore, identifying a marker is expected to address the problem by providing an objective measurement.

The main reason for focusing on speech is that asking potential depression patients to speak is something that can be done easily in a clinical setting. This is a major advantage with respect to biological markers investigated so far that require invasive exams like, e.g., brain neurotrophic factors (Lee et al., 2007) or monoamine levels in cerebrospinal fluids (Placidi et al., 2001). Not to mention that these markers have been investigated while developing pharmacological treatments and have been shown to be affected by several drawbacks, including the difficulty

to interpret placebo-controlled trials (H. Yang et al., 2005; Fava et al., 2003) methodological pitfalls at the level of patient selection and enrolment (Demitrack et al., 1998) or misalignment between clinicians' observations and patients' self-assessments (R. P. Greenberg et al., 1992). Similarly, the efforts of the computing community have explored a wide spectrum of behavioural markers (facial expressions, nonverbal vocal behaviour, etc.), but none of them appears to clearly outperform the others. Furthermore, compared to the analysis of read speech, other behavioural cues might be difficult to capture and analyse outside a laboratory setting.

For these reasons, this chapter investigates the use of computational paralinguistics (Schuller & Batliner, 2013) and social signal processing (Vinciarelli et al., 2009) as a means to identify effective depression markers in speech, i.e., observable and measurable traces of the pathology in the way people speak. The markers in this chapter account for two main behaviours, namely speech duration and the use of silences. The main reason for focusing on such behaviours is that, according to neuroscience, one of the main effects of depression is that the brain tends to become slower at processing linguistic information. Therefore, it is reasonable to expect that depressed individuals tend to take longer time in speaking and to be less fluent. Compared to the previous work Esposito et al. (2016), this work converts behaviour differences into detectable speech representations in the model in an automatic way.

The experiments of this chapter involved 118 participants included in the Androids Corpus (see Section 3.2). The markers are examined over two types of speech, i.e., the read speech of 112 participants (the maximum number of participants performing a Read Task and the spontaneous speech of 116 participants (the maximum number of participants performing an Interview Task), see Section 3.2). The approach based on a standard feature set, originally designed to recognise emotions (Schuller et al., 2009), has been used to perform initial experiments. The feature set has then been expanded with the markers, and the difference in performance resulting from the expansion of the feature set has been used as a measure of how effectively the behaviours above can account for the presence of depression. The results show that adding 3 features to the initial 32 is sufficient to increase the accuracy from 69.7% to 82.1% for the read speech and from 71.5% to 85.8% for the spontaneous speech, corresponding to a reduction of the error rate by 40.9% and 50.2%, respectively. This is of particular interest when considering that the new features are less than 10% of the original set (3 out of 32). Furthermore, the results appear to be in line with the findings of neuroscience about brain-level differences between depressed and non-depressed individuals. In other words, reading speed and silences appear to be reliable markers of depression in different types of speech.

The rest of this chapter is organised as follows: Section 4.2 surveys previous work, Section 4.3 describes the data used in the experiments, Section 4.4 reports on experiments and results, while Sections 4.5 draws some conclusions.

4.2 Previous Work

The computing community has made substantial efforts towards the automatic detection of depression in speech (see (Cummins, Scherer, et al., 2015) for an extensive survey). In general, researchers have tackled two primary tasks in this area. The first task is the inference of scores obtained through the administration of self-assessment questionnaires (e.g., the Beck Depression Inventory II). The second task is actual depression detection, i.e., automatic discrimination between people diagnosed with depression by psychiatrists and control individuals that are not affected by mental health issues. In a few cases, the efforts have targeted the identification of depression markers, as exemplified in this thesis.

In several cases, the proposed approaches have taken into account not only the speech signals but also their corresponding transcriptions (Morales & Levitan, 2016; Williamson et al., 2016; Rohanian et al., 2019; Al Hanai et al., 2018). In particular, Morales & Levitan (2016) specifically maintain that it is crucial to consider both acoustic and linguistic aspects of speech when developing depression-related technologies, as opposed to addressing them separately. However, findings from other studies, such as those by Williamson et al. (2016) have indicated that the most optimal outcomes are achieved when using transcriptions only. Furthermore, experiments conducted by Rohanian et al. (2019) suggest that the multimodal combination of paralinguistic features and text yields effective results only when deep networks with attention gates are employed. While these observations may not universally apply and are influenced by the limitations of the models and the characteristics of the datasets, they have demonstrated effectiveness within the context of their respective experiments. Finally, in the case of the work by Al Hanai et al. (2018) the authors propose that the best results can be obtained by taking into account the interaction dynamics, which involve considering the context in which a sentence is uttered during a conversation.

Overall, the collective findings from the above studies do not provide a clear conclusion on whether considering what people say is helpful or not in the context of depression detection. However, the commonality of those studies is that their focus is on the sole speech signal like in a large number of other contributions including, e.g., (Huang et al., 2018; L.-S. A. Low et al., 2010; Cummins, Sethu, et al., 2015). In the first work (Huang et al., 2018), the experiments aim at testing whether speech samples captured through mobile phones allow one to discriminate between people that are above or below a threshold score of the Personal Health Questionnaire. The results show that this is actually possible with an accuracy of 72%. In the other two works (L.-S. A. Low et al., 2010; Cummins, Sethu, et al., 2015), the goal is to identify depression markers that, like in this chapter, can help to distinguish between depressed individuals and the others. The focus of the experiments in (L.-S. A. Low et al., 2010) is on adolescents because their voice is not fully formed and, therefore, speech-based depression detection can be a more challenging task. The results of the work (Schuller & Batliner, 2013) show that the most effective marker is the energy of the signal, corresponding to how loud people speak, especially

when measured with the Teager operator. In contrast, Cummins, Sethu, et al. (2015) identified the variability of phonetic characteristics as the most effective marker for the detection of depression.

In this chapter, we conduct experiments on speech signals, following the approach of numerous preceding studies (see above). Although there is no clear consensus regarding the specific type of speech to be analyzed, our experiments encompass both read and spontaneous speech samples. In this respect, this allows to evaluate the effectiveness and robustness of the proposed markers across diverse speech contexts in distinguishing between depressed and non-depressed speakers.

4.3 The Data

In this chapter, we carry out a series of experiments that involve 118 participants in the Androids Corpus. These experiments incorporate two different types of speech: read speech, which involves 112 participants (Read Task), and spontaneous speech, which includes 116 participants (Interview Task). For a more detailed account of the corpus and the specific tasks undertaken in this research, please refer to Section 3.2.

4.4 Experiments and Results

In this chapter, we present an approach and experimental design that follow the same protocol as the baseline, which is comprehensively discussed in Section 3.3. One of the key aspects of this process involves the conversion of speech signals into feature vectors through a feature extraction step, as elaborated in Section 3.3.1. It is important to note that the size of analysis windows and time steps employed in this chapter are consistent with those used in the baseline approaches, as detailed in Section 3.3.2.

The dimensionality of each generated feature vector is 32, which includes not only the 16 core features (as outlined in Section 3.3.1), but also their delta coefficients. The latter represents the differences between the features in the current analysis window and their corresponding values in the previous window.

At the end of the feature extraction step, we calculate the average of the feature vectors and combine them with the proposed markers in this Chapter (discussed in further detail below). This combined data is then fed into the Support Vector Machine (SVM) for recognition, utilizing the same parameters as those employed in the baseline approaches described in Section 3.3.2. The recognition is performed using the k-fold protocol (k = 3) and adheres to the same data split as BS1, as presented in Section 3.3.4.

Additionally, it is worth mentioning that the model training is conducted on the same device and relies on the same version of the scikit-learn library as used in BS1. This consistency ensures

a fair comparison and is in line with the information provided in Section 3.3.4.

4.4.1 Duration of Speech

Neuroscience research has provided valuable insights into the relationship between language processing and depression, suggesting that some brain processes revolving around language tend to take more time in people affected by depression. In particular, it has been shown that there is an association between depression and dysfunctions in several areas, such as the frontal gyrus and Pre-Frontal Cortex (PFC), involved in semantic language processing (Seghier et al., 2004).

Furthermore, it has been observed that when interpreting the meaning of words, depression patients display slower activation of the left temporoparietal (Wernicke) area, a key region for language comprehension, and involvement of brain regions (including right lateral PFC) not activated in the case of non-depressed people (Abdullaev et al., 2002). These findings suggest that the neural mechanisms underlying language processing in depressed individuals might be different from those in healthy individuals.

As a consequence of these altered neural processes, individuals with depression require more time to assign meaning to words. In this respect, it is reasonable to expect that depressed participants take more time to read the text and answer the same questions at the core of the experiments. In other words, it is reasonable to expect that the amount of time needed to read the text and answer the same questions can act as a depression marker.

In an analysis presented in Section 3.2, the average time and standard deviation for reading the text are 52.9 ± 10.9 and 47.4 ± 8.8 seconds for depressed and control participants, respectively. Given that all participants read a total of 185 words, this corresponds to average reading speeds of 209.8 and 234.2 words per minute, respectively, for depressed and control participants. These results are consistent with the neuroscience indications discussed earlier, further supporting the hypothesis that language processing is affected in individuals with depression and can leave traces in speech.

In addition, the tendency of depressed patients to require more time to speak is also observed in the other types of speech, i.e., spontaneous speech. Given the statistically significant difference observed in the duration of read speech (as reported in Section 3.2) between depressed and healthy people, it is possible to expect that the use of the duration of speech as a feature, in addition to the 32 features of the standard set, should lead to an improvement of the performance.

The results of such an intervention are summarized in Table 4.1. When compared to the baseline classifier, the accuracy improved by 6.2 points for read speech and 5.9 points for spontaneous speech, corresponding to error rate reductions of 20.5% and 16.7%, respectively. According to two-tailed binomial tests, these improvements are statistically significant ($p < 0.05$ for both instances), providing further evidence for the effectiveness of using speech duration as a feature in depression detection. In summary, the time an individual takes to read a given text or

Table 4.1: Performance after taking the duration of speech into account. RT (Read Task) and IT (Interview Task) are the results for read and spontaneous speech, respectively.

Clssifier	Accuracy (%) (%)	Precision (%)	Recall (%)	F1-Score (%)
RT	75.9	75.1	79.9	77.2
IT	70.6	75.5	70.9	72.2

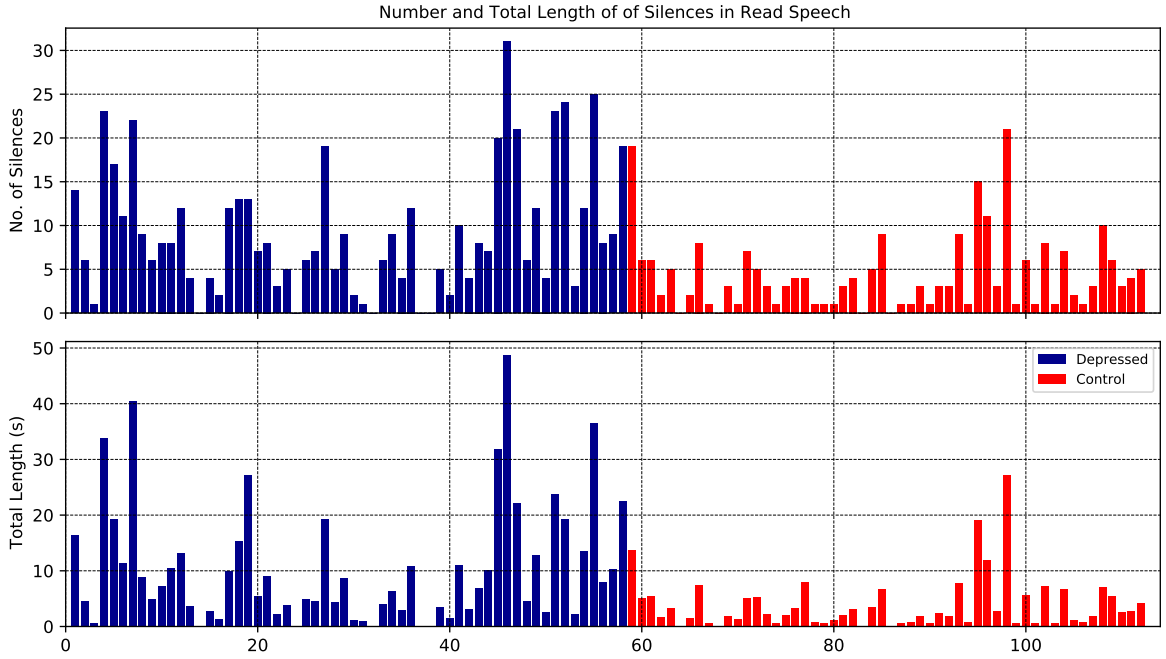

Figure 4.1: Silences distribution across participants in read speech (Read Task). The top chart shows the number of silences per participant, the lower one shows the average length per participant.

respond to questions spontaneously appears to be a promising and effective depression marker that could potentially improve diagnostic and intervention strategies.

4.4.2 Effect of Silences

The previous section shows that brain-level differences between depressed and non-depressed individuals lead to observable differences in the amount of time needed for speaking. Furthermore, the differences are consistent enough to significantly improve the accuracy of a baseline classifier in both types of speech through the addition of just one feature to the original set of 32. This section shows that it is possible to further improve the performance of the baseline classifier by taking into account how depression changes the neural processes responsible for verbal fluency. In fact, the literature shows that, in the brain of depressed people, the median prefrontal cortex and angular gyrus generate interferences that result in speech disfluency, typ-

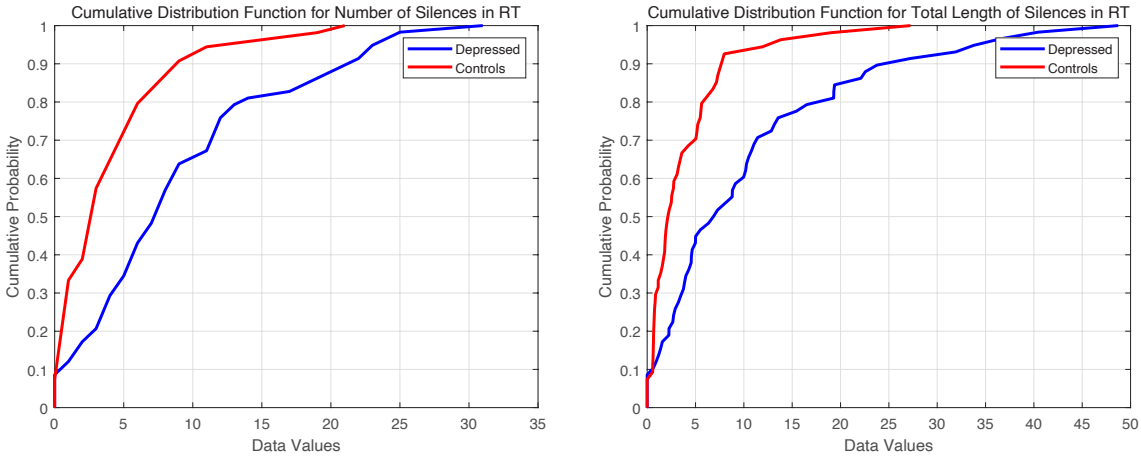

Figure 4.2: Cumulative distribution function for silences in read speech (Read Task).

ically through the recruitment of a larger number of brain areas involved in speech initiation and higher-order language processes (Backes et al., 2014). Not surprisingly, many studies found that depressed individuals have deficits in phonemic and verbal fluency (Wagner et al., 2012). Furthermore, improved fluency is typically used as a signal of melioration during depression treatment (Wagner et al., 2012).

One possible effect of the differences above is that depressed people tend to display more frequently intervals of time during which there is no emission of voice. Furthermore, for depressed people, these intervals of time might tend to be longer. For this reason, the voicing probability in the baseline feature set has been used to identify sequences of consecutive frames in which voice is unlikely to be emitted. This has led to the estimate of the probability $p(r)$ of r consecutive frames to show a null voicing probability. Correspondingly, it led to the identification of a minimum threshold value r_0 such that the following holds:

$$p(r \geqslant r_0) = \sum_{r=r_0}^{r_{max}} p(r) \leqslant 0.05, \tag{4.1}$$

where r_{max} is the maximum value of r observed in the data. Given the number of participants is different for Read Task and Interview Task in this chapter (not every participant performs both tasks), r_0 is estimated for read and spontaneous speech separately.

In the experiments of this chapter, r_0 is 55 for read speech and 72 for spontaneous speech, corresponding to a length of 0.565 s and 0.735 s, respectively, not far from the conventional threshold of 0.5 s used to identify pauses in linguistics (Schuller & Batliner, 2013). Hereafter, sequences of null voicing probability at least r_0 frames long are referred to as *silences*. Figure 4.1 and Figure 4.3 show the distribution of the number of silences across participants in read and spontaneous speech. Figure 4.2 and Figure 4.4 show the corresponding cumulative distribution function for silences. The average number of silences for depressed and control participants in read speech is 9.3 and 4.3, respectively ($p < 0.001$ according to a two-tailed t-test). For the

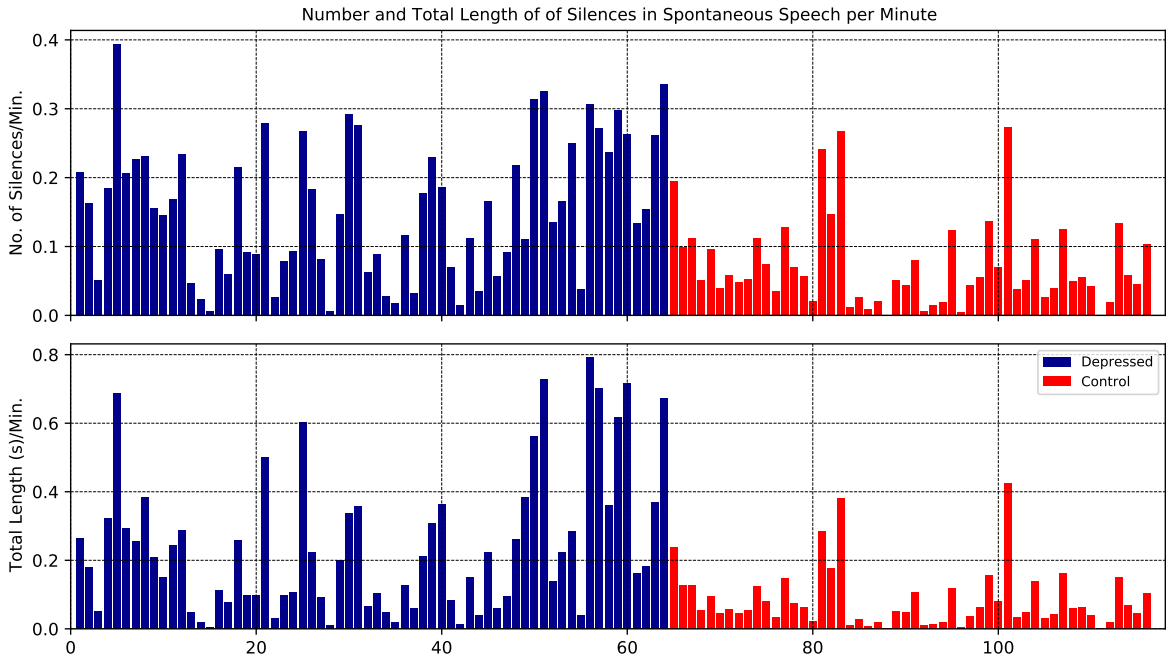

Figure 4.3: Silences distribution across participants in spontaneous speech (Interview Task). The top chart shows the number of silences per participant, the lower one shows the average length per participant.

Table 4.2: Performance after taking both duration and silences into account. RT (Read Task) and IT (Interview Task) are the results for read and spontaneous speech, respectively.

Clssifier	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
RT	82.1	80.3	86.6	83.2
IT	82.7	84.3	85.5	84.2

spontaneous speech, the average number of silences for depressed and control participants is 29.3 and 15.1, respectively ($p < 0.01$ according to a two-tailed t-test). When it comes to the total length of silences in the read speech, it is 10.6 s and 3.9 s for the two groups. Such a difference is statically significant according to a two-tailed t-test ($p < 0.001$). In the cases of spontaneous speech, the total length of silences is 45.2 s for depressed and 17.4 s for control participants. Furthermore, such a difference is once again statistically significant according to a two-tailed t-test ($p < 0.001$). These results suggest depressed people tend to speak with more and longer pauses in both types of speech. This reveals that the feature set can be expanded with the number and total length of silences. In other words, these two features might act as depression markers.

The recognition results are reported in Table 4.2. Compared to the results obtained after adding the length of the recordings to the original set of 32 features, there is a further increase

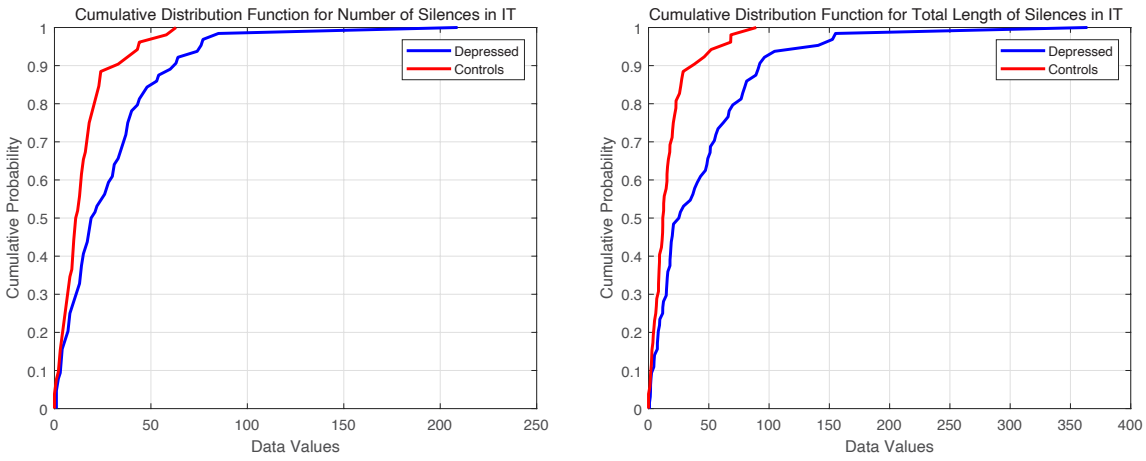

Figure 4.4: Cumulative distribution function for silences in spontaneous speech (Interview Task).

by 6.2% points of the accuracy, corresponding to a further reduction by 25.7% of the error rate in the read speech. Such an improvement is statistically significant with $p < 0.05$ according to a binomial test). In the cases of spontaneous speech, there is also a further increase by 12.1% of the accuracy, corresponding to a further reduction by 41.2% of the error rate. Such an improvement is also statistically significant with $p < 0.001$ according to a binomial test). Besides increasing the performance of the approach, the two features described in this section provide a possible explanation of why depressed participants tend to take more time to speak. In particular, the effectiveness of the two features suggests that depressed individuals do not just read slower, they tend to spend more time without uttering the words they read, possibly because they need more time to process the linguistic information involved. For spontaneous speech, depressed people show the same tendency as in read speech. However, compared with the use of speech duration, the use of silences becomes more effective in spontaneous speech (with 12.0% improvement of F1-Score) than in read speech (6.0% improvement of F1-Score). This may result from the greater proportion of silences in spontaneous speech, especially given the statistically significant reduction in the duration of spontaneous speech in depressed patients (see Section 3.2). These results confirm that silences can be used as markers in automatic depression detection.

4.5 Conclusions

This chapter has presented experiments aimed at answering the hypothesis proposed in the thesis statement that depression leaves traces in the temporal aspect of speech. Identifying these depression markers in speech, i.e., of measurable speech characteristics that can help to distinguish between depressed and non-depressed individuals. Compared to most previous works in the literature, the methodology used to identify the markers is not based on statistical testing or correlational analysis, but on the performance improvement observed when using the

markers as features in a classifier. In particular, the experiments show that three features accounting for three different markers (duration of speech, number of silences and total length of silences) increase the accuracy from 69.7% to 82.1% for read speech and from 64.7% to 82.7% for spontaneous speech when added to an initial set of 32 features (statistically significant with $p < 0.0001$ according to a two-tailed binomial test). These latter were selected as a baseline because, while originally designed to capture emotion (Schuller et al., 2009), are known to effectively account for a much wider spectrum of psychological phenomena, including depression (see Section 3.3.1).

In addition, compared to most previous work in the literature, this chapter has tried to combine computational paralinguistics (Schuller & Batliner, 2013), based on low-level speech features extracted from 25 ms long windows, and social signal processing (Vinciarelli et al., 2009), based on the detection of nonverbal behavioural cues associated to a phenomenon of interest. This latter aspect is important because markers should correspond to observable aspects of behaviour, given that they must be of help for the diagnosis of depression. In other words, compared to measurements like fundamental frequency or MFCCs, the advantage of observable behaviours like reading speed or use of silences is that they can be possibly observed without the need for automatic analysis.

The experiments have been performed over read and spontaneous speech, i.e., over recordings of people asked to read the same text and answer the same series of questions. The reason behind the choice is to examine the effectiveness of proposed markers in the two different types of speech, i.e., read and spontaneous speech. In other words, the use of read speech limits the effect of variability sources not necessarily related to depression while the use of spontaneous speech introduces more variability due to the cognitive effort in planning what to say next. The results in this chapter confirm the effectiveness of the markers and the robustness when applying the markers to different types of speech. This is one of the reasons why the markers appear to be in line with the indications of neuroscience showing that depressed people tend to take more time to process linguistic information and to be more disfluent.

One interesting aspect of the markers considered in the work is that they are likely to be honest (Pentland, 2010), i.e., sufficiently difficult to control consciously to allow one to fake them. For example, in the read speech, the silence length differences between depressed and control participants correspond to an average silence length of 0.32 s and 0.28 s, respectively. Similarly, the speed differences correspond to an average time per read word of 286 ms and 256 ms for depressed and control participants, respectively. Both differences are too subtle to be consciously controlled. Therefore, it is unlikely that a depression patient can try to look like a non-depressed individual. This is an important advantage because people affected by mental health issues can try to hide their condition in order to escape treatment, mainly to avoid the stigma associated to psychiatric problems. In this respect, the markers proposed in this chapter promise to be of help for clinicians dealing with potential depression patients.

In this chapter, we have concentrated on the markers in the temporal aspect of speech for depression detection, while not exploring potential markers in the acoustic aspect, which is a more relevant dimension of speech. Moving forward to the next chapter, we intend to address this limitation by proposing and examining potential markers that are specifically associated with the acoustic aspect of speech.

Chapter 5

Feature Correlation Matrices for Depression Detection

5.1 Introduction

In Chapter 4, we introduced timing-based depression markers in speech, such as speech duration and silences. These markers support our thesis statement that depressed individuals exhibit measurable speech characteristics that can help distinguishing between depressed and non-depressed individuals. The promising performance of our proposed timing-based speech markers motivates us to further advance their development. However, these timing-based speech markers are an indirect measurement of depression because they do not reflect the changes in acoustic properties found in depressed speech. In fact, studies suggest that numerous acoustic features change during speech production among depressed speakers, as discussed in Chapter 2. This raises the following question: *Do the relationships between acoustic features also change in the speech of depressed speakers, and can such changes serve as potential markers for depression detection?*

Multiple approaches use speech as input because it has been demonstrated that depression leaves traces in the way people speak (see Section 5.2). In fact, according to neuroscience, brain connectivity patterns tend to be more unstable in depression patients (Wise et al., 2017). Specifically, depressed speakers tend to exhibit a lower degree of coordination across different brain areas, which can "[...] alter speech production by influencing the characteristics of the vocal source, tract, and prosodics [and lead to] psychomotor retardation, where a patient shows sluggishness and motor disorder in vocal articulation, affecting coordination across multiple aspects of production" (Williamson et al., 2013). Such a phenomenon is expected to change not only the acoustic properties of speech but also the relationships between them. Consequently, it is anticipated to alter the correlations between features and how these correlations change over time.

Despite this, to the best of our knowledge, the literature pays only limited attention to the

relationships between features (see Section 5.2). The key assumption of this chapter is that this oversight results in missing the major depression effects described earlier. For this reason, the focus of this work is on investigating whether feature correlations and their changes over time can help improve the effectiveness of depression detection.

The experiments were performed over the Androids Corpus, using an available public dataset, and were firstly performed on read speech for two main reasons. First, the brain phenomena mentioned earlier were also observed when people read (Regev et al., 2013). Second, read speech involves less variability resulting from factors not necessarily related to the pathology (e.g., the topic being discussed). Then the proposed approach was applied to spontaneous speech to examine the effectiveness of the approach on different types of speech for depression detection. The speech samples were represented as sequences of feature correlation matrices (see Section 5.4), and this representation was shown to improve two classification approaches, namely Support Vector Machines (SVM) and Long Short-Term Memory (LSTM) networks (Hochreiter & Schmidhuber, 1997). Compared to feature vectors, correlation matrices reduced the error rate by up to 29.0% and up to 15.1% for SVMs and LSTMs, respectively, in read speech. When it comes to spontaneous speech, the use of correlation matrices reduces the error rate by up to 36.5% and up to 11.4%. Overall, at least for the architectures tested, the proposed approaches achieved an F1 score of up to 86.0% and 85.9% when fed with feature correlation matrices in read speech and spontaneous speech, respectively. Additionally, using sequences of feature correlation matrices made the classification process approximately twice as fast.

In summary, the main contributions of this chapter are as follows:

- To the best of our knowledge, this is the first work comparing the performance of the same models (SVM and LSTM) when fed with feature vectors and with feature correlation matrices;
- This is the first work demonstrating that changes in correlation patterns over time can act as a depression marker.

Comparisons with previous results obtained using the same data show that the above novelties lead to higher performances, on average.

The rest of this chapter is organized as follows: Section 5.2 describes related work, Section 5.3 describes the data, Section 5.4 explains the approach used in the experiments, Section 5.5 reports on experiments and results, and the final Section 5.6 draws some conclusions.

5.2 Previous Work

In recent years, the computing community has made substantial efforts towards automatic depression detection, with the identification of depression markers in speech being one of the most important areas of focus (Cummins, Scherer, et al., 2015; Wu et al., 2022; Highland &

Zhou, 2022). Researchers have observed that speech production differs in depressed individuals (Cummins, Scherer, et al., 2015), and several studies have attempted to identify markers related to temporal properties. For instance, speech rate (Morales & Levitan, 2016) and frequency of pauses (Tao et al., 2020) have been found to significantly improve depression detection, resulting in a 15-point increase in accuracy when classifying depressed and non-depressed speakers. However, the effectiveness of these markers is shown to vary depending on the type of speech (Kiss & Vicsi, 2017a).

Instead of identifying markers, many approaches represent speech in terms of feature vectors and employ various machine learning methodologies to detect depression. Low-Level Descriptors are the primary features used for this purpose, typically extracted from short analysis windows. The standard practice involves concatenating these speech features with their time derivatives (delta) to account for potential temporal effects. Such features have proven to be robust in characterizing both read and spontaneous depressive speech (Cummins et al., 2011; Cummins, Scherer, et al., 2015; Alghowinem, Goecke, Wagner, Epps, Breakspear, & Parker, 2013). Mel Frequency Cepstral Coefficients (MFCCs) are the most common features, often enhancing performance (Taguchi et al., 2018; Niu et al., 2019; Rejaibi et al., 2022). Researchers have also found that fusing MFCCs with other speech features (such as F0) can further improve depression detection performance (Wu et al., 2022). However, these typical speech features, including MFCCs, are sensitive to channel variability (Stasak & Epps, 2017) and can be easily affected by noise, making them less than ideal (Mitra et al., 2016).

The proposed approach was based on measuring the correlation between features extracted at multiple time distances, resulting in multiple correlation matrices concatenated with each other (Williamson et al., 2013). This correlation structure successfully reflects coordination information between formant frequencies and delta-mel-cepstrum in depression detection. In contrast to Low-Level Descriptors, the correlation structure provides information on intricate variations between feature domains. This approach has proven effective not only for depression but also for epileptic seizures (Williamson et al., 2011). Furthermore, it has been demonstrated that this correlation structure can effectively handle variable-length data in feature extraction (Dong & Yang, 2021). More recent studies show that such a correlation representation is effective with deep neural networks as well (Song et al., 2020; Huang, Epps, & Joachim, 2020). However, there are two main issues associated with the aforementioned correlation structure. The first issue is that it focuses on time-delayed information, necessitating longer speech recordings (the time distances for measuring the correlation can be high). Consequently, this can lead to reduced performance when dealing with short-duration speech utterances (Williamson et al., 2016). The second main issue is the appearance of artefacts that require extra steps in preprocessing (Huang et al., 2019a). The approach presented in this chapter addresses both of these issues by measuring correlations over shorter time intervals and modeling changes in correlations using sequential models. To the best of our knowledge, there is limited knowledge

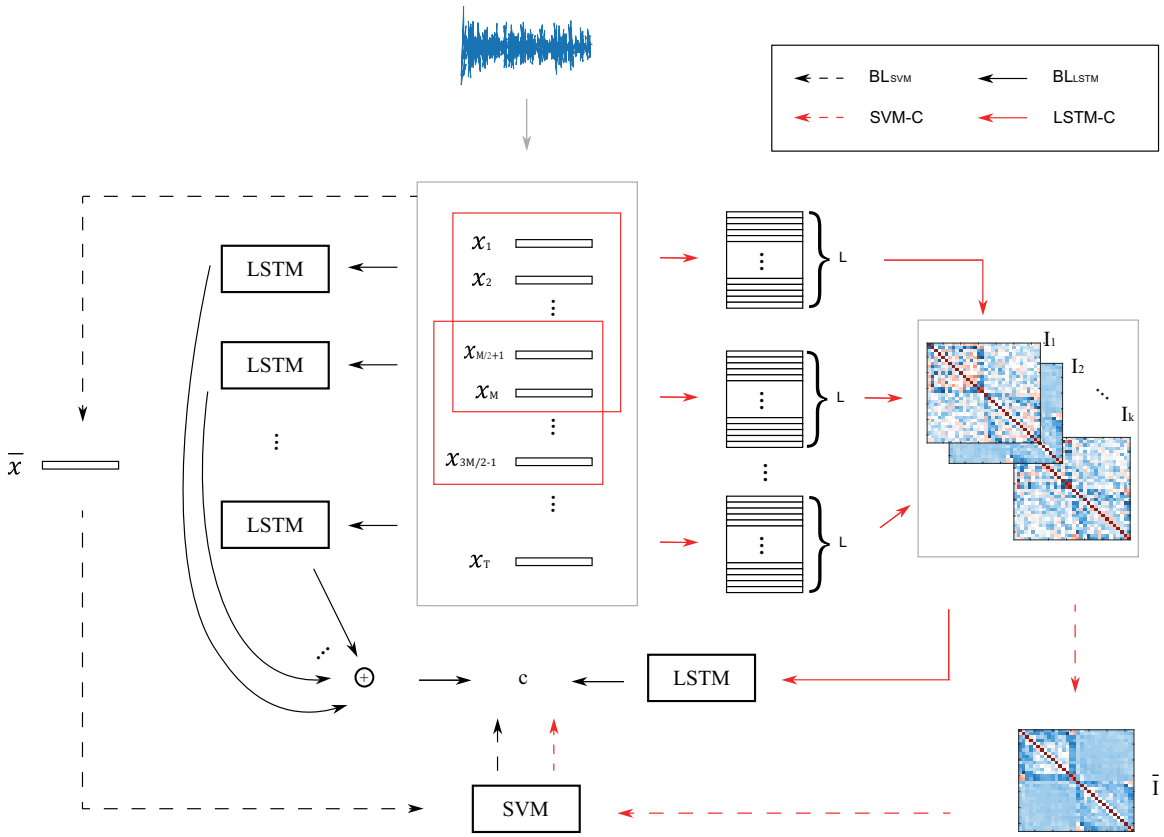

Figure 5.1: The diagram shows the four approaches used in the experiments. The black arrows stand for the use of feature vectors while the red arrows stand for the use of correlation matrices. The black dashed arrows stand for BL_{SVM}, while the black arrows stand for $SVM-C$. The red dashed arrows stand for BL_{LSTM}, while the red arrows stand for $LSTM-C$. The symbol $\oplus$ corresponds to aggregation.

about the dynamic changes of correlation structures over time, and the integration of correlation structure with acoustical speech features has not been explored before.

5.3 The Data

This chapter included a total of 118 participants from the Androids Corpus. The Read Task involved recordings from 112 participants, while the Interview Task involved recordings from 116 participants. For more specific details, please refer to Chapter 3, Section 3.2.

5.4 The Approach

The goal of the experiments is to test whether feature correlation matrices convey more depression-relevant information than feature vectors. For this reason, the approach (see Figure 5.1) includes a *feature extraction* step (conversion of speech signals into sequences of feature vectors), a *cor-*

relation representation step (mapping of sequences of feature vectors into sequences of feature correlation matrices), a *depression detection* step and a *stability measurement* step. Figure 5.1 shows that this latter can be performed by feeding the same model (SVM or LSTM) with either feature vectors or correlation matrices. In this way, it is possible to test whether these latter convey more information than the sole features.

5.4.1 Feature Extraction

The aim of the feature extraction is to convert the speech recordings into sequences of feature vectors $\vec{x}_k$. The extraction in this chapter is the same as the step in Chapter 3 Section 3.3.1 including the same feature set, size of analysis windows as well as tools for extraction. This leads to the same dimension of feature vectors ($D = 32$) as those used in the baselines (BL_{SVM} and BL_{LSTM}) in Chapter 3.

5.4.2 Correlation Representation

The correlation representation step segments the sequence of feature vectors into subsequences of length M (the number of vectors included in one subsequence) starting at regular steps of length $M/2$ (two consecutive subsequences overlap by half of their elements). In the experiments, the value of M ranges between 100 and 500, corresponding to time intervals of length between 1 and 5 seconds. Once the segmentation is performed, it is possible to extract a *local feature correlation matrix* W_k as follows:

$$\{W_k\}_{ij} = corr(X_i, X_j), \tag{5.1}$$

$$X_i = \{\vec{x_k}\}_i, \tag{5.2}$$

$$X_j = \{\vec{x_k}\}_j, \tag{5.3}$$

where $k \in [1, N]$, N is the total number of subsequences, X_i and X_j are the values of feature i and j along the sequence of vectors X, *corr* is the function of Spearman Correlation (Howell, 2009), element $\{W_k\}_{ij}$ is the correlation coefficient between X_i and X_j ($i, j \in [1, D]$, D is the number of features) in the correlation matrix. The reason for using such a correlation is that the feature values extracted from the feature extraction step do not follow the normal distribution. Given that there are multiple subsequences, there will be multiple local feature correlation matrices W_k. The start and the end of vectors (Q_s and Q_e, respectively) in the subsequence corresponding to the matrix W_n are as follows:

$$Q_s = (n-1)M/2, \tag{5.4}$$

$$Q_e = [(n-1)/2+1]M - 1 \tag{5.5}$$

where $n = 1, \ldots, N$. All correlation coefficients are converted into Z-scores with Fisher's transformation, a standard step in statistical analysis of dependent correlations (Meng et al., 1992).

5.4.3 Depression Detection

The first baseline approach for depression detection is BL_{SVM}, which takes the average of the feature vectors extracted from a recording and feeds it to a linear kernel SVM (see Chapter 3 for more details). Similarly, given that the correlation representation step converts every speech signal into a sequence of matrices W_k, it is possible to estimate the average W_k and feed it to a Support Vector Machine to perform the depression detection step (see $SVM-C$ in Figure 5.1). The correlation matrices are symmetric and only the elements below the principal diagonal are used, thus leading to a dimension as follows:

$$\hat{D} = D(D-1)/2 = 496 \tag{5.6}$$

where $D = 32$ is the original number of features. BL_{SVM} and $SVM-C$ can be compared to test whether feature correlation matrices are actually of help.

Another baseline to perform depression detection is BL_{LSTM} which splits the sequence of the feature vectors into subsequences including 128 consecutive vectors (a standard value in the literature) and then feed each one of these to LSTMs, (see Chapter 3 for more details). Once all subsequences are classified, it is possible to apply a majority vote and assign a recording to the class its subsequences are more frequently assigned to (BL_{LSTM} in Figure 5.1). The same approach can be applied to the sequence of the W_k. In this case, a linear layer reducing the dimension of the correlation matrices to 32 is added to the LSTM with the goal of making the problem computationally more tractable ($LSTM-C$ in Figure 5.1). BL_{LSTM} and $LSTM-C$ can be compared to test whether the matrices actually improve over the feature vectors.

5.4.4 Stability Measurement

The goal of this step is to estimate a measurement for the changes of correlation matrices over time, namely *stability*. The reason for investigating such changes is that, as shown in Figure 5.2, the *local* correlation matrices (one of the correlation matrices in the sequences) provide more variation information in depressed and control groups when compared with the *global* correlation matrices (the average of correlation matrices for two groups). This suggests that the use of local correlation matrices provides a possible way to discriminate between depressed and non-depressed speakers.

After the correlation representation step, the original speech signals are converted into the sequence of correlation matrices W_k ($k \in \{1, \ldots, N\}$). This allows to define a measurement of

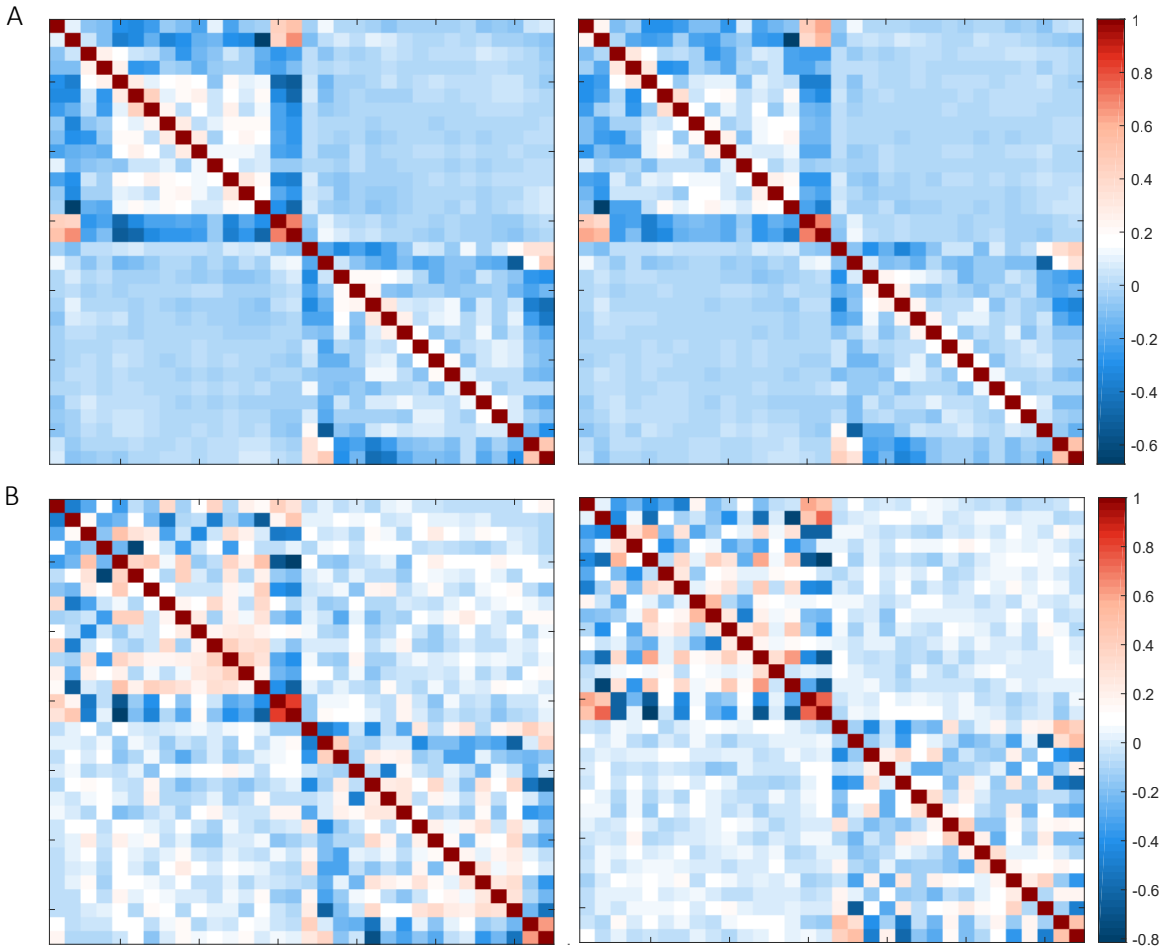

Figure 5.2: Figure A shows the average of W_k for the control (left) and depressed groups (right), namely *global* correlation matrices. Figure B shows one of specific W_k for the control (left) and depressed groups (right), namely *local* correlation matrices.

the stability $\hat{S}$ of local correlation matrices as follows:

$$\hat{S} = \frac{1}{N}\sum_{1}^{N} corr(W_k, W_{k+1}), \tag{5.7}$$

where W_k and W_{k+1} are the consecutive correlation matrices in the sequences, $k \in \{1,\ldots,N\}$, N is the total number of correlation matrices, *corr* is the function of Spearman Correlation (Howell, 2009). Given that the number of consecutive vectors M ranges between 100 and 500 in a frame, this lead to different W_k at different M. Therefore, the stability $\hat{S}$ will be dependent on the value of M. Due to discontinuity in spontaneous speech (each original X in spontaneous speech is split into several sub-sequences to discard the voice of the interviewer), stability in spontaneous speech is estimated as follows:

$$\hat{S} = \frac{1}{s}\frac{1}{N}\sum_{1}^{s}\sum_{1}^{N} corr(W_k, W_{k+1}), \tag{5.8}$$

where W_k and W_{k+1} are the consecutive correlation matrices in the subsequence, $k \in \{1,\ldots,N\}$,

Table 5.1: Depression detection results in terms of Accuracy, Precision, Recall and F1 score. The approaches are numbered 1 to 4 according to Figure 5.1. The performance metrics are represented in terms of average and standard deviation obtained over 10 repetitions of the experiment. At every repetition, the weights of the LSTM were initialized to different random values. The table reports the best accuracy over different lengths L. Suffix C means classifier with correlation matrices.

	Accuracy	Precision	Recall	F1 Score
Reading Task				
Random	50.1	51.8	51.8	51.8
BL_{SVM}	69.7±6.6	69.6±3.5	76.3±12.9	71.9±4.8
SVM-C	78.5±4.0	78.5±8.8	78.8±3.2	78.5±6.0
BL_{LSTM}	83.4±2.6	85.0±3.5	83.0±2.3	83.8±2.1
LSTM-C	85.9±1.3	85.5±2.5	87.1±2.7	86.0±1.1
Interview Task				
Random	50.5	55.2	55.2	55.2
BL_{SVM}	64.7±6.3	68.4±7.5	66.3±10.9	66.6±7.1
SVM-C	77.6±3.3	82.5±13.7	77.8±6.9	78.9±3.5
BL_{svm}	81.6±1.6	83.3±2.7	85.3±1.4	83.4±1.1
LSTM-C	83.7±0.4	81.9±1.3	90.6±2.4	85.9±0.4

N is the total number of correlation matrices in the subsequence, *corr* is the function of Spearman Correlation (Howell, 2009), s is the number of subsequences from one speaker.

5.5 Experiments and Results

The experiments were performed according to the same k-fold protocol ($k = 3$) as in the BL_{SVM} and BL_{LSTM}, see Chapter 3 Section 3.3 for more details about experimental design. The folds were disjoint and were the same as the data split in BL_{SVM} and BL_{LSTM}. The same participant never appeared in both training and test set and this ensures that the approach actually detects depression and does not simply recognize the speaker. All hyperparameters were set a-priori and no attempt was made to find configurations possibly leading to higher performances.

The number of hidden states in the LSTMs was set to $H = 32$ (the same number of hidden states as the baselines in the public corpus). The number of training epochs was set to 100 and the learning rate was set to 0.0005. The training was performed using the RMSProp as an optimizer (Tieleman et al., 2012) and the categorical cross-entropy as a loss function. All experiments were replicated $R = 10$ times using a different initialization of the networks and data splitting at each repetition. For this reason, all results are reported in terms of average and standard deviation observed over the repetitions. This ensures that the performances are not the result of a favorable initialization, but a realistic estimate of the system's effectiveness.

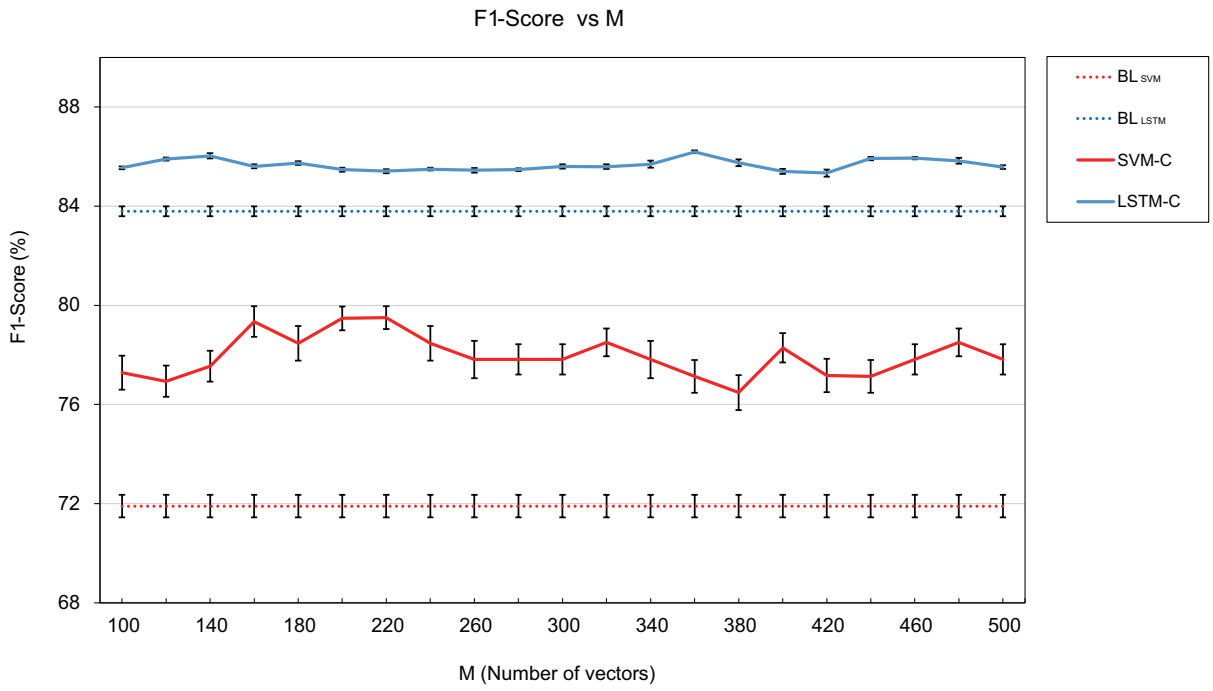

Figure 5.3: The figure shows averaged F1-Score and its standard error of the mean over 10 repetitions at different lengths M in read speech. C means classifier with correlation matrices.

The goal of these experiments is to compare the performance of the same model when using feature vectors or feature correlation matrices as input. For this reason, Table 5.1 shows the results in terms of Accuracy, Precision, Recall and F1 Score. The results of random baseline approach, BL_{SVM} and BL_{LSTM} are the same as in Table 3.4 and Table 3.5. The table shows the performance of SVMs and LSTMs when giving as input both the sequence of vectors X and the sequence of feature correlation matrices W_k. According to a two-tailed t-test, all approaches perform better than the random baseline ($p < 0.001$ in all cases). Furthermore, always according to two-tailed t-tests, models perform better when fed with feature correlation matrices than when fed with feature vectors ($p < 0.05$ for all cases). The error rate decreases by 29.0% when passing from BL_{SVM} to $SVM-C$ in read speech, and the relative reduction of the error rate is up to 15.1% for approaches based on LSTMs in read speech. Similarly, the error rate decreases by 36.5% and up to 11.4% when passing from BL_{SVM} to $SVM-C$ and passing from BL_{LSTM} to $LSTM-C$, respectively. In this respect, the results suggest that feature correlation matrices actually convey more depression-relevant information than simple feature vectors.

Given BL_{SVM} and BL_{LSTM} lack the correlation representation step, the results of these approaches are irrelevant to the selection of M, shown as BL_{SVM} and BL_{LSTM} in Figure 5.3 and Figure 5.4. Table 5.1 shows the best results across all values of parameter M. For this reason, Figure 5.3 and Figure 5.4 show how the F1-Score changes depending on M in read and spontaneous speech, respectively. The results suggest that the models fed with feature correlation matrices tend to outperform those fed with feature vectors, thus confirming that the results of

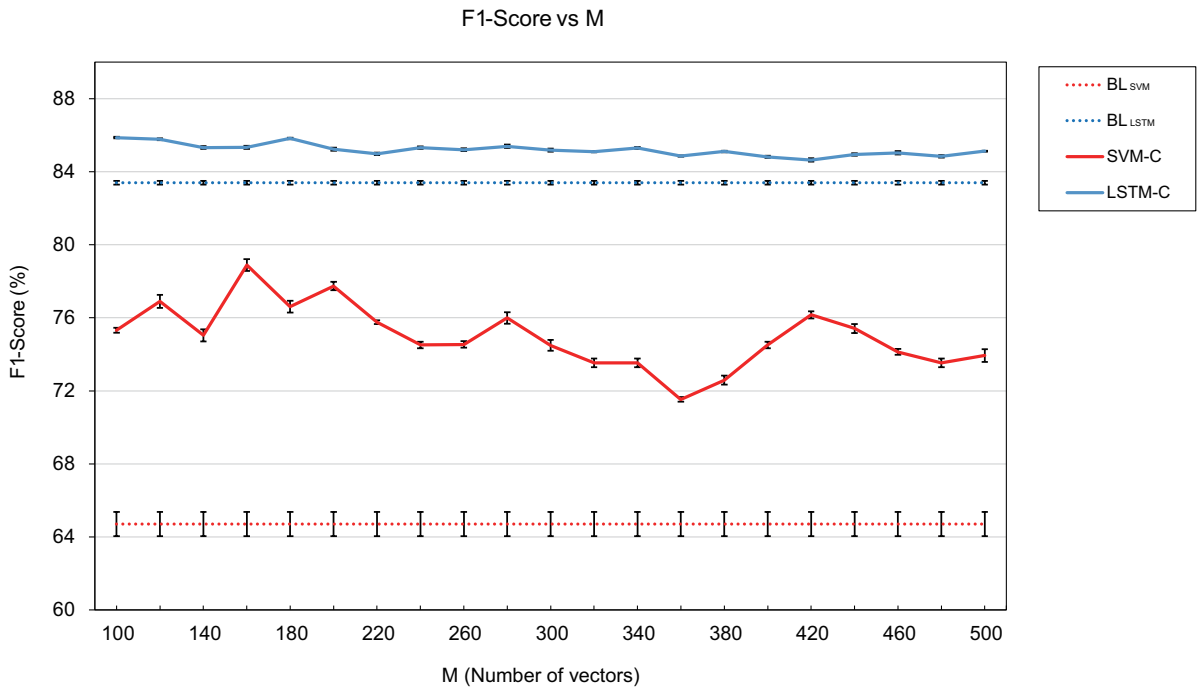

Figure 5.4: The figure shows averaged F1-Score and its standard error of the mean over 10 repetitions at different lengths M in spontaneous speech. C means classifier with correlation matrices.

Table 5.1 are not the result of a particular choice of M. Furthermore, these results also suggest that the improvement by using correlation matrices in depression detection is irrelevant to the choice of speech type, meaning the use of correlation matrices is robust in different types of speech and can actually capture depression-related information.

In addition, the use of the feature correlation matrices reduces significantly the amount of time required to train the LSTMs. When using the same computing infrastructure (Google CoLab Tesla T4 GPU), the time goes from 4.5 to 2 minutes for training LSTMs, such an improvement is statistically significant according to a two-tailed t-test ($p < 0.001$), corresponding to a 56% reduction of time use. This is mainly because LSTMs need fewer cells in the case of correlation matrix sequences, see Figure 5.1. In the case of SVMs, no significant difference was observed in terms of training time. A further advantage of correlation matrix sequences is that, being shorter, they do not need to be segmented into subsequences to be fed to the LSTMs. Therefore, it is possible to avoid the majority vote.

One possible explanation of the results above is in Figure 5.5. The chart shows an example of the distribution of stability when $M = 300$. This suggests that the distribution of stability for the depressed and control group tend to be different. Figure 5.6 shows the average of stability over different values of M. A two-tailed t-test shows that, for all values of M, the stability is higher, to a statistically significant extent, in the case of control participants ($p < 0.05$ with FDR correction for all values of M) in both read and spontaneous speech. This result suggests the speech of depressed patients involves more variability and tends to be more unstable. In

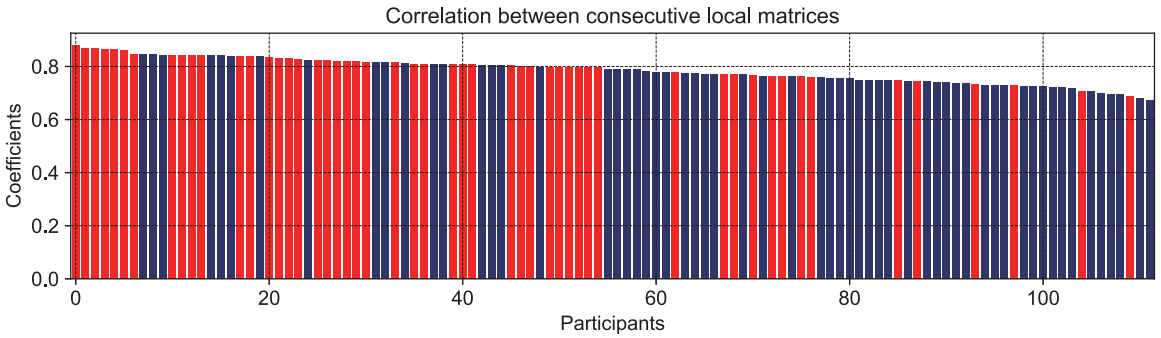

Figure 5.5: The chart shows an example of the distribution of stability for control (red bars) and depressed (blue bars) participants, ordered from highest to lowest. Control speakers tend to be more frequent in the first part of the chart and this suggests that the correlation tends to be higher for them.

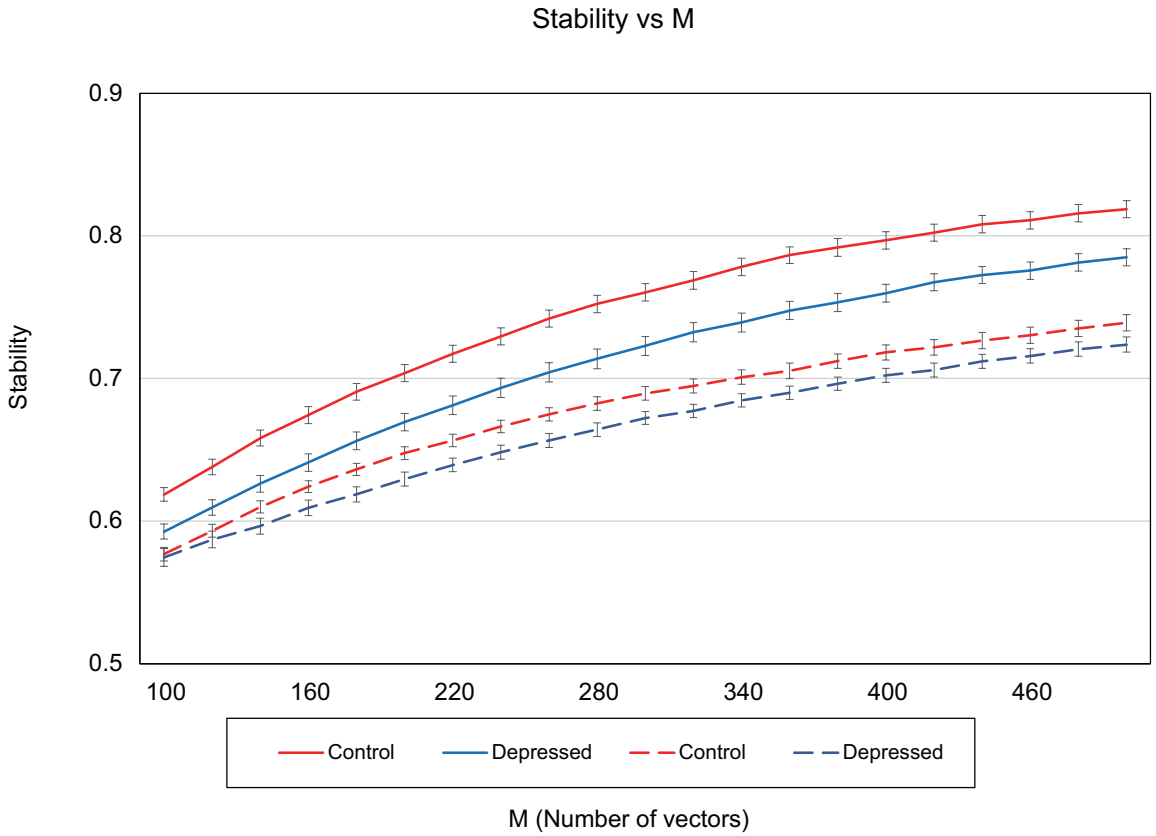

Figure 5.6: The chart shows the average of stability and their standard error of the mean over the different values of M. The red line and the red dashed line stand for the average stability of control in read speech and spontaneous speech, respectively. The blue line and the blue dashed line stand for the average stability of depressed patients in read speech and spontaneous speech, respectively.

particular, the correlation is around 0.75 for control participants and around 0.71 for depressed ones, for all values of M. Such an observation suggests that the relationship between features extracted at different points in time tends to be less consistent in the case of depressed people

Table 5.2: Performance in terms of Accuracy, Precision, Recall and F1-Score after taking the stability into account in the BL_{SVM}. RT (Read Task) and IT (Interview Task) are the results for read and spontaneous speech, respectively.

	RT				IT			
M	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
100	78.6	80.4	78.5	79.0	74.1	78.8	74.6	75.4
120	77.7	78.8	78.5	78.3	73.3	77.6	74.6	74.8
140	79.5	81.9	78.5	79.6	75.0	79.9	74.5	76.2
160	78.6	80.4	78.5	79.0	74.1	80.5	73.2	75.1
180	78.6	82.0	77.1	78.9	74.1	78.5	75.8	75.8
200	80.4	82.0	80.5	81.0	74.1	78.8	74.5	75.5
220	76.8	78.5	76.6	77.4	71.5	75.4	74.5	73.7
240	77.7	79.9	76.6	78.0	72.4	76.5	74.4	74.3
260	75.1	75.8	76.6	76.1	73.3	78.1	74.5	74.9
280	76.8	75.6	81.3	78.3	70.7	75.5	71.1	72.1
300	78.6	79.5	79.9	79.6	68.9	73.5	71.1	71.0
320	74.2	74.0	77.9	75.9	66.4	71.1	68.4	68.3
340	75.1	75.4	77.9	76.6	63.8	67.3	68.4	66.7
360	78.6	78.1	83.8	80.2	64.7	68.7	67.0	67.0
380	76.0	75.3	80.5	77.6	63.8	67.8	67.0	66.4
400	76.9	75.1	83.8	79.1	61.2	65.0	65.6	64.3
420	77.7	76.3	83.8	79.5	63.0	67.2	65.6	65.3
440	78.7	78.4	82.4	80.2	63.8	68.7	65.6	65.8
460	76.8	75.3	83.8	78.9	62.1	65.9	65.6	64.9
480	76.8	76.7	80.5	78.0	62.1	65.9	65.6	64.9
500	76.8	76.9	80.5	78.1	62.1	66.3	65.6	64.7

and this is probably a depression marker that helps the models to perform better.

According to a two-tailed t-test, stability in read speech is statistically significantly higher than those in spontaneous speech for both groups ($p < 0.05$ with FDR correction for all values of M). One possible explanation for these findings is that spontaneous speech includes different speech contents, which results in increased variability compared to read speech. Consequently, this greater variability leads to higher instability in spontaneous speech.

In addition, a two-way analysis of variance (ANOVA) reveals a significant interaction effect ($p < 0.05$) between the type of speech (i.e., read or spontaneous speech) and condition (i.e., depression or control). The goal of using ANOVA is to examine whether the relationship between the type of speech and condition is independent. In this chapter, the interaction effect is significant, indicating that the effect of the type of speech on stability differs between the depressed and non-depressed groups. In other words, the difference in stability between the two groups is more pronounced in the read speech compared to the spontaneous speech.

Another advantage of stability is that it allows to covert correlation matrices into Low-Level

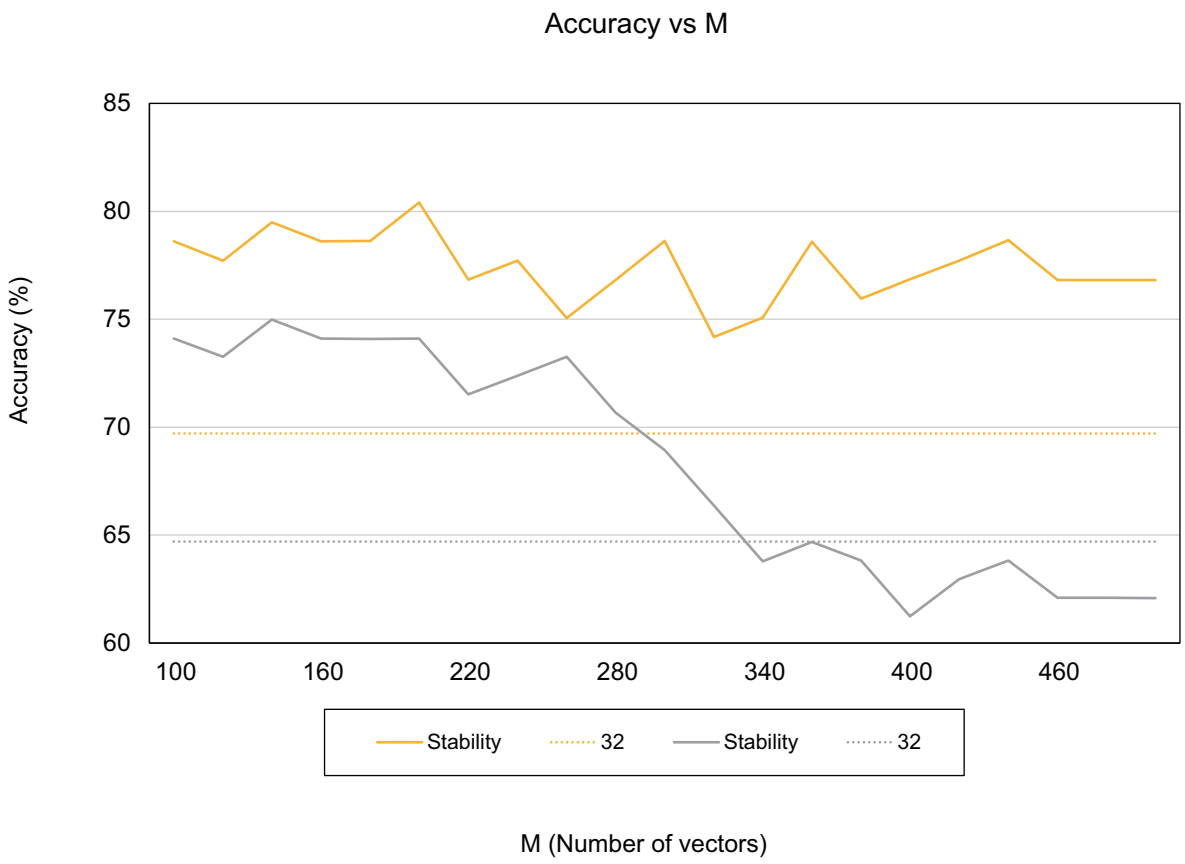

Figure 5.7: The chart shows the accuracy when adding stability to the original 32 features in BL_{SVM} over different M. The yellow line and the grey line stand for using 32 features and stability in BL_{SVM} in the read and spontaneous speech, respectively. The yellow dotted line and the grey dotted line stand for only using 32 features in BL_{SVM} in the read and spontaneous speech, respectively.

Descriptors. In this respect, it provides a possible way to combine typical speech features and correlation matrices together in depression detection. Table 5.2 presents the recognition results after adding stability to the original set of 32 features in the read and spontaneous speech. According to a two-tailed t-test, for all values of M, adding stability to the original 32 feature set results in a statistically significant ($p < 0.05$ for all values of M) improvement in accuracy for read speech, as shown in Figure 5.7. The improvement between the best results and BL_{SVM} in read speech reaches up to 10.7%, corresponding to the reduction of the error rate by 35.3%. When it comes to spontaneous speech, there is also a further increase by 10.3% of the accuracy (as indicated by the best accuracy in Table 5.6), corresponding to a further reduction by 29.2% of the error rate. However, such an improvement is only significant ($p < 0.05$) in the smaller values of M ($M < 320$). These results confirm the previously identified interaction effect between the type of speech and the condition, indicating that stability tends to be more beneficial for read speech compared to spontaneous speech.

Table 5.3: Previous studies involving the same participants.

Studies	Accuracy	Precision	Recall	F1
Scibelli et al. (2018)	77.0	74.0	80.0	77.0
Tao et al. (2020)	84.5	84.5	84.6	84.5
Aloshban et al. (2020)	83.5	95.0	70.3	80.5
Aloshban et al. (2021)	83.0	95.2	69.0	80.0
Aloshban et al. (2022)	84.7	**95.4**	72.4	82.3
Alsarrani et al. (2022)	67.6	71.7	72.2	72.3
Ours	**85.9**	85.5	**87.1**	**86.0**

5.6 Conclusions

In this chapter, we answer the research questions proposed in the introduction. The key-result of this chapter is that feature correlation matrices lead to better depression results than feature vectors in both read and spontaneous speech, at least in the case of linear kernel SVMs and LSTMs, the two models used in the experiments. The proposed *stability* further explains such an improvement and confirms the effectiveness of using correlation matrices. In addition, the proposed end-to-end approach significantly reduces the time for detection. To the best of our knowledge, this is the first work that compares feature vectors and correlation matrices in terms of the performance they lead to. Furthermore, this is the first work proposing an explanation of the observed results in terms of a possible marker (the different correlation between consecutive matrices). It seems not possible to directly integrate the feature correlation matrices with the markers identified in Chapter 4 for depression detection. The reason for this is that silences serve as global descriptors, whereas feature correlation matrices function as local descriptors.

Table 5.3 shows results obtained in other works involving the speakers from the same public dataset as those considered in this chapter. Our results achieve the highest performance outperforming state-of-the-art approaches and baselines. In comparison to previous studies (Aloshban et al., 2020, 2021, 2022), our proposed approach demonstrates a distinct trade-off between precision and recall. While these studies exhibit higher precision, their recall is significantly lower than that of our approach. Notably, our approach outperforms the others in terms of the F1-score, a crucial metric that balances both precision and recall, thereby highlighting the effectiveness of the proposed methodology. The aim of this chapter is, to examine the effectiveness of correlation matrices by comparing them with feature vectors, but never involved with the combination of both directly. Therefore, we will introduce the possibility in the combination of two types of speech for depression detection in the later chapters.

The main limitation of the experiments is the use of a linear layer in $LSTM-C$. Its aim is to keep the same input dimensionality as BL_{LSTM}, while still preserving the correlation between the features. This costs extra-parameters that make it less clear, in the comparison between BL_{LSTM}

and $LSTM-C$, whether the performance improvement actually results from the correlation matrices. On the other hand, in the case of the SVMs, the only change between BL_{SVM} and $SVM-C$ is the use of the matrices and the improvement is statistically significant. This seems to confirm that the matrices actually help to improve depression detection. In the next chapter, we aim to tackle such an issue by proposing a novel deep learning model without extra-parameters. Then we will focus on combining different types of speech for depression detection.

Chapter 6

Multi-local Attention for Depression Detection

6.1 Introduction

In the previous chapters, we explored speech markers beneficial for automatic depression detection, supporting the hypothesis in our thesis statement that depressed individuals exhibit traces in their speech that can be used for identifying depression. However, it is crucial to acknowledge that these chapters only discussed potential markers advantageous for depression detection without suggesting specific improvements to the model. The computing community is currently making significant strides towards developing technologies that capture valuable information for automatic depression detection more effectively. This leads to the following question: *can deep learning methods enhance depression detection performance without increasing the model parameters?*

As demonstrated in Chapter 3, Support Vector Machines (SVMs) can achieve an F1-Score of up to 69.7% when provided with the average of all feature vectors from a speech recording, surpassing the baseline performance on the same dataset used in this study. Such an observation implies that feature vectors that are in proximity to the average convey information relevant to depression and should, therefore, be considered more reliable. In this respect, recent work (Linsley et al., 2019) suggests that attention mechanisms can be employed to highlight feature vectors that are likely to contain information related to a given task.

Several deep learning approaches have been proposed for automatic depression detection including convolutional neural networks (CNNs) (He & Cao, 2018; A. Y. Kim et al., 2023; Rodrigues Makiuchi et al., 2019) and recurrent neural networks (RNNs) (Niu et al., 2019; Rejaibi et al., 2022; Aloshban et al., 2022). In particular, the use of attention mechanisms has become increasingly prevalent in the deep learning community. The attention mechanisms (Vaswani et al., 2017) include group of algorithms in the context of deep learning, which aims to emphasize the most useful features or information related to the task being performed. The application of

attention mechanisms has become successful in a broad range of fields including speech recognition (Yeh et al., 2019), image retrieval (Ge et al., 2023), video description (Zhu & Jiang, 2019) etc, and more recently, in automatic depression detection (Z. Zhao et al., 2019; Niu et al., 2020; Yin et al., 2023).

On one hand, attention mechanisms are useful for processing multi-modal data, as they enable the extraction of different types of information from various modalities to improve the overall representation of task-relevant information. Recent studies (Ray et al., 2019; Qureshi et al., 2019; Niu et al., 2020; Rodrigues Makiuchi et al., 2019) have proposed using attention mechanisms to extract complementary information between modalities for enhancing the multi-modal representation of depression for classification purposes. Conversely, applying attention mechanisms to single-modal data, such as speech, is not only effective but also requires less computational power compared to processing multi-modal data. Additionally, this approach offers practical benefits for diagnosis, including simpler data collection. For the acoustics aspect, the authors C. Cai et al. (2021) introduced attention-based CNNs focusing on time-domain speech signals to extract multi-scale contextual information related to depression, enabling effective classification. Regarding linguistics, the authors (Z. Zhao et al., 2019) proposed a hierarchical attention network emphasizing depression-related linguistic information. These studies suggest that attention mechanisms can effectively emphasize task-relevant information across various data modalities.

In this chapter, we introduce a novel approach, namely *Multi-Local Attention*, an attention mechanism (Vaswani et al., 2017) that enhances depression detection based on Long Short-Term Memory Networks (LSTMs) (Hochreiter & Schmidhuber, 1997). The goal of this approach is to emphasize feature vectors more similar to the local average within the corresponding *frame*, a brief segment extracted from a speech recording. The rationale behind using a local average is its ability to provide a more accurate representation of potential changes over time. In fact, earlier studies employing attention mechanisms for the same task relied on incorporating attention layers in Deep Networks, an established but relatively older approach, see, e.g., (Harati et al., 2021; Z. Zhao et al., 2020; Aloshban et al., 2022). The experiments were performed on read and spontaneous speech, in line with a large body of previous research, see, e.g., (Mitra & Shriberg, 2015; Alghowinem, Goecke, Wagner, Epps, Breakspear, & Parker, 2013; Liu, Li, et al., 2017; Tao et al., 2020). The main advantage of using two types of speech is to provide a comparison of the model applied to different data types.

In total, the experiments in this work involved 118 participants, including 64 diagnosed with depression by professional psychiatrists, who were asked to perform Read and Interview Tasks (refer to Chapter 3 for a detailed description of the tasks). The results showed that the Multi-Local Attention (MLA) approach improved the performance of depression detection based on LSTMs. Furthermore, most importantly, the MLA proposed in this work increases the accuracy of baselines from 73.3% to 88.1% for the Read Task, corresponding to increases from 74.7%

to 88.5% in terms of F1-Score, thus reducing the error rate by 44.6%. When it comes to the Interview Task, MLA increases the accuracy of baselines from 64.1% to 86.0% for read speech, corresponding to increases from 71.6% to 87.4% in terms of F1-Score, thus reducing the error rate by 44.4%. Moreover, the improvement in performance was achieved without a significant increase in the number of model parameters, demonstrating the potential of the Multi-Local Attention model for automatic depression detection. According to the results, Multi-Local Attention is beneficial not only for enhancing performance metrics such as Accuracy and F1-Score, but also for improving two other critical aspects of the approach, both important from an application point of view. The first aspect pertains to the effectiveness of a confidence score associated with the detection outcome in identifying speakers who are more likely to be correctly classified. The second aspect reduces the amount of data required to classify a speaker as depressed or non-depressed.

The contributions of this chapter can be summarized as follows:

- To the best of our knowledge, this is the first attempt to use the Multi-Local Attention approach for depression detection.
- We present a comprehensive evaluation of the proposed model, examining its performance from various perspectives such as performance analysis, effectiveness assessment, time efficiency analysis, and comparative performance across diverse speech types.

The rest of this chapter is organized as follows: Section 6.2 describes the data, Section 6.3 describes the approach used in the experiments, Section 6.4 reports on experiments and results, and the final Section 6.5 draws some conclusions.

6.2 The Data

In this chapter, we use a total of 118 participants, of which 116 participants have the data of read speech (RT Task) and 112 participants have the data of spontaneous speech (IT Task). This allows for a thorough performance evaluation of our proposed model across different types of speech. For more details related to the data, please refer to the description in Chapter 3.

6.3 The Approach

The approach used in the experiments includes four main steps, namely *feature extraction*, *multi-local attention*, *recognition* and *aggregation* (see Figure 6.1). It is important to note that although the following descriptions are specifically related to read speech, these steps can be easily adapted and applied to spontaneous speech as well. All approaches were trained and tested separately over the Read and Interview Tasks.

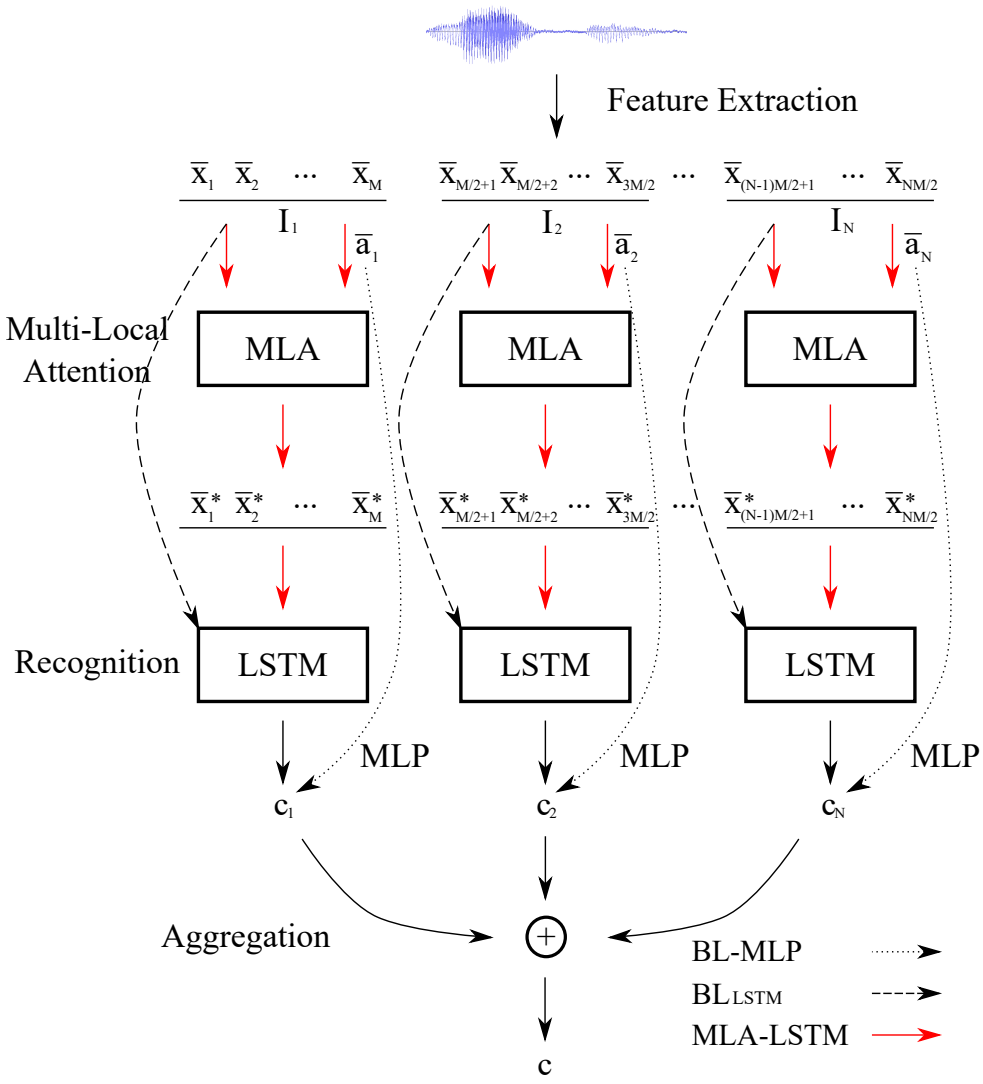

Figure 6.1: The figure shows the main steps of the approach. Vectors $\vec{a}_k$ are the averages extracted from every frame, MLA stands for Multi-Local Attention, c_k is the classification outcome for frame I_k, the symbol $\oplus$ corresponds to the majority vote and c is the final classification outcome.

6.3.1 Feature Extraction

In the first step, feature extraction, we align with the procedure outlined in Chapter 3, using the same feature set and feature extraction approach. This consistency is crucial to ensure a fair comparison between the proposed approach and baseline methods. The feature extraction step effectively processes each recording and represents it as a sequence $X = \{\vec{x}_1, \ldots, \vec{x}_T\}$ of T feature vectors. This process remains consistent with the baselines described in Chapter 3 Section 3.3.

Following the completion of the feature extraction process, the sequences are divided into frames $I = \{I_k\}$ ($k \in \{1, \ldots, N\}$) to facilitate further analysis. Each frame, labeled as I_k, includes $M = 128$ consecutive vectors aligning with the BL_{LSTM} approach (refer to Section 3.3 for more

information). To maintain consistency with the BL_{LSTM} method, the frames also start at regular steps of length $M/2$, which ensures that the overlap size between two consecutive frames remains identical to that of the BL_{LSTM} approach.

6.3.2 Multi-Local Attention

After the segmentation into frames, BL_{LSTM} (the same as shown in Figure 3.2 in Chapter 3) and BL-MLP move to the recognition step (see dashed and dotted arrows in Figure 6.1). Meanwhile, the MLA takes a different approach and performs the *Multi-Local Attention* step (see red arrows in Figure 6.1).

In the first stage of the *Multi-Local Attention* step, the algorithm calculates the average feature vector, denoted by $\vec{a}_k$, for each *k-th* frame. Subsequently, it applies a transformation to the remaining feature vectors within the same frame according to the following equations:

$$\vec{x}_i^* = \vec{x}_i + \cos(\theta_i) \cdot \vec{x}_i, \tag{6.1}$$

$$\cos\theta_i = \frac{\vec{a}_k \vec{x}_i}{||\vec{a}_k|| ||\vec{x}_i||}, \tag{6.2}$$

where $\cos\theta_i$ is the cosine of the angle between $\vec{a}_k$ and $\vec{x}_i$, $i \in \{1,\ldots,M\}$, ($\cos\theta_i$ is typically referred to as *cosine similarity*). The transform emphasizes the vectors that are more closely aligned with $\vec{a}_k$ by increasing their norm. As a result, a feature vector orthogonal or opposite to the average will be mapped into a null vector, while the average itself will see its norm multiplied by $\sqrt{2}$. Any other vector will be between such extremes, effectively highlighting their relevance to the average vector.

6.3.3 Recognition

After the MLA step, the proposed MLA method proceeds to the recognition phase, where it feeds the transformed vectors into LSTMs. In comparison, the BL-MLP performs the recognition by feeding the averages extracted from the frames to a Multi-Layer Perceptron (see dotted arrows in Figure 6.1), while the BL_{LSTM} performs it by feeding the vectors of a frame to LSTMs. In all cases, the frame is assigned either to class *depressed* or to class *control*. It is important to note that both BL_{LSTM} and BL-MLP skip the MLA step before performing the recognition.

6.3.4 Aggregation

Given that there are multiple frames per recording, there is an *aggregation* step to take every individual frame classification outcome to a majority vote for all approaches. In this process, the recording is assigned to the class its frames are most frequently assigned to. This aggregation step is consistent with the procedure described in Chapter 3. One of the main advantages of this

approach is its ability to facilitate a strict comparison between the proposed MLA and baselines. Another advantage of employing this aggregation technique is that it is possible to define a confidence score, represented by s, according to the following equation:

$$s = \frac{n(\hat{c})}{N}, \tag{6.3}$$

where $n(\hat{c})$ is the number of frames assigned to the majority class, which corresponds to the class that is most frequently associated with the frames in a given recording. The main assumption behind such a definition is that the tendency to assign a greater fraction of frames to a given class should be associated to correct classification results.

6.4 Experiments and Results

The experiments were performed according to the same k-fold cross-validation protocol ($k = 3$) as the baseline approaches to ensure a consistent evaluation framework, as detailed in Section 3.3 of Chapter 3. This procedure facilitates a fair and unbiased comparison between the MLA and the established baseline approaches.

In the experimental setup, the LSTMs were configured with 32 hidden states. The learning rate was set to 10^{-3} to control the update step size during the training process, and the training was carried out for 300 epochs to allow the model to converge. The RMSProp optimizer (Tieleman et al., 2012) was employed for training, utilizing categorical cross-entropy as the loss function to measure the discrepancy between the predicted and true labels (Rubinstein & Kroese, 2004).

Given that LSTMs require a random initialization for their weights, the experiments were repeated $R = 10$ times to account for potential variability in the results. In each repetition, the weights were initialized differently. For such a reason, all performance metrics are reported in terms of average and standard deviation across the R repetitions to provide a comprehensive overview of the model performance.

Both LSTM-based versions of the approach (BL_{LSTM} and MLA) employed the same parameters and training procedures to maintain consistency in the comparison. The MLP, another type of neural network used in the study, featured 32 hidden states. The training process for the MLP was the same as that of the LSTMs to ensure a fair comparison. MLA and BL-MLP approaches were implemented with the same PyTorch version as BL_{LSTM} (version in 1.13.1+cu116) (Paszke et al., 2019) to guarantee compatibility and reproducibility of the experimental results.

6.4.1 Performance Analysis

Table 6.1 presents a comprehensive overview of the recognition results for all approaches in terms of Accuracy, Precision, Recall and F1-Score for both Read and Interview Tasks. All

Table 6.1: Recognition results in terms of Accuracy, Precision, Recall and F1-Score. R and I stand for Read and Interview Task, respectively.

	Task	Acc.	Prec.	Rec.	F1
BL-MLP	R	73.3±1.6	73.5±2.4	78.0±3.0	74.7±1.0
BL_{LSTM}	R	85.5±1.2	84.1±2.6	86.8±2.3	84.7±1.3
MLA	R	**88.1±1.6**	**87.5±2.7**	**90.0±2.6**	**88.5±1.3**
BL-MLP	I	64.1±1.3	64.2±1.8	82.7±4.4	71.6±1.0
BL_{LSTM}	I	83.1±0.8	80.4±2.1	87.8±2.7	83.3±0.7
MLA	I	**86.0±1.3**	**87.3±4.0**	**88.2±3.9**	**87.4±1.0**

approaches were compared to a random classifier, as described in Equation 3.7 in Chapter 3 Section 3.3. Precision, Recall and F1-Score correspond to the prior of the positive class (i.e., *depression* in these experiments) which is 51.8% for the Read Task and 55.2% for the Interview Task. According to a two-tailed t-test, all approaches outperform the random classifier to a statistically significant extent ($p < 0.001$ in all cases after Bonferroni correction). These results suggest that all the approaches in this chapter are effective at recognizing depression in both Read and Interview Tasks.

The only difference between BL_{LSTM} and BL-MLP is the use of LSTM for recognition. The use of LSTM increases the accuracy by 12.2% and 19.0% for Read and Interview Tasks, respectively. This enhancement leads to a substantial reduction in the error rate, decreasing by 45.7% for the Read Task and by 52.9% for the Interview Task. This substantial improvement suggests that LSTM is highly effective at boosting recognition performance in the context of depression detection. The observed improvements are statistically significant, with $p < 0.001$ according to two-tailed t-tests, applied with the False Discovery Rate (FDR) correction.

Furthermore, the only difference between the BL_{LSTM} and MLA approaches is the application of the Multi-Local Attention mechanism before feeding data into the LSTM. This additional step further increases the accuracy by 2.5% for the Read Task and 2.9% for the Interview Task, corresponding to a reduction of the error rate by 17.2% for both data types. These improvements are also statistically significant, with $p < 0.001$ according to two-tailed t-tests (after applying the FDR correction). This result indicates that the MLA effectively emphasizes depression-relevant information in each data type, and therefore enhancing the performance of the model.

An analysis of the results reveals that the performance of all approaches for the Read Task surpasses that of the Interview Task. This observation is in line with previous studies in the field, see e.g., (Alghowinem, Goecke, Wagner, Epps, Breakspear, & Parker, 2013; Kiss & Vicsi, 2017a), which report a performance difference of approximately 3.0% between read and spontaneous speech. In addition, Table 6.2 demonstrates that the MLA approach achieves results comparable to those obtained in previous studies that involve the same speakers considered in this work. Overall, the results appear to support to the key assumption behind the proposed

Table 6.2: Recognition results in terms of Accuracy, Precision, Recall and F1 Score. The table includes the results obtained in this chapter and the results from previous studies involving the same dataset. R and I stand for Read and Interview Task, respectively (I+Text means that both Interview Task and its transcription were used).

	Task	Acc.	Prec.	Rec.	F1
Tao et al. (2020)	R	84.5	84.5	84.6	84.5
Scibelli et al. (2018)	R	77.0	74.0	80.0	77.0
Aloshban et al. (2020)	I+T	83.5	95.0	70.3	80.5
Aloshban et al. (2021)	I+T	83.0	95.2	69.0	80.0
Aloshban et al. (2022)	I+T	84.7	**95.4**	72.4	82.3
Alsarrani et al. (2022)	I	67.6	71.7	72.2	72.3
Ours	R	**88.1±1.6**	87.5±2.7	**90.0±2.6**	**88.5±1.3**
Ours	I	**86.0±1.3**	87.3±4.0	**88.2±3.9**	**87.4±1.0**

Multi-Local Attention mechanism: vectors closer to the local average should be "trusted" more, as they likely contain more relevant information for the task at hand.

6.4.2 Confidence Score Analysis

Section 6.3 shows that the proposed approach associates a confidence score s to its classification outcomes, providing a useful metric for evaluating the model performance. By considering only the participants with the r highest values of s, in other words, the top r ranked speakers in terms of the confidence score, we can evaluate the effectiveness of the approach in distinguishing speakers who are more likely to be classified correctly.

Figure 6.2 shows how the accuracy changes as a function of r (the vertical bars correspond to the standard error of the mean over the R repetitions of the experiment) for both Read and Interview Tasks. The curves show that the application of the Multi-Local Attention mechanism leads to notable improvements in accuracy at every point in the ranking for both tasks. This result is statistically significant, with $p < 0.001$ according to a two-tailed t-test. This finding suggests that the Multi-Local Attention enhances the effectiveness of the confidence score, making it a more reliable metric for identifying speakers who are more likely to be classified correctly for both types of speech.

Moreover, by analyzing the data presented in Figure 6.2, it becomes evident that it is possible to set a threshold value h. When considering only participants with confidence scores of $s \geq h$, the accuracy corresponds to a predefined value. For example, when using the confidence value at rank $r = 23$ as a threshold in the Interview Task, the accuracy reaches an impressive 99%, which notably outperforms the other two values of 90% and 58%, respectively. These results seem to confirm that the confidence measure tends to be greater in correspondence with correct classification outcomes. In this way, this insight allows for a more efficient allocation of

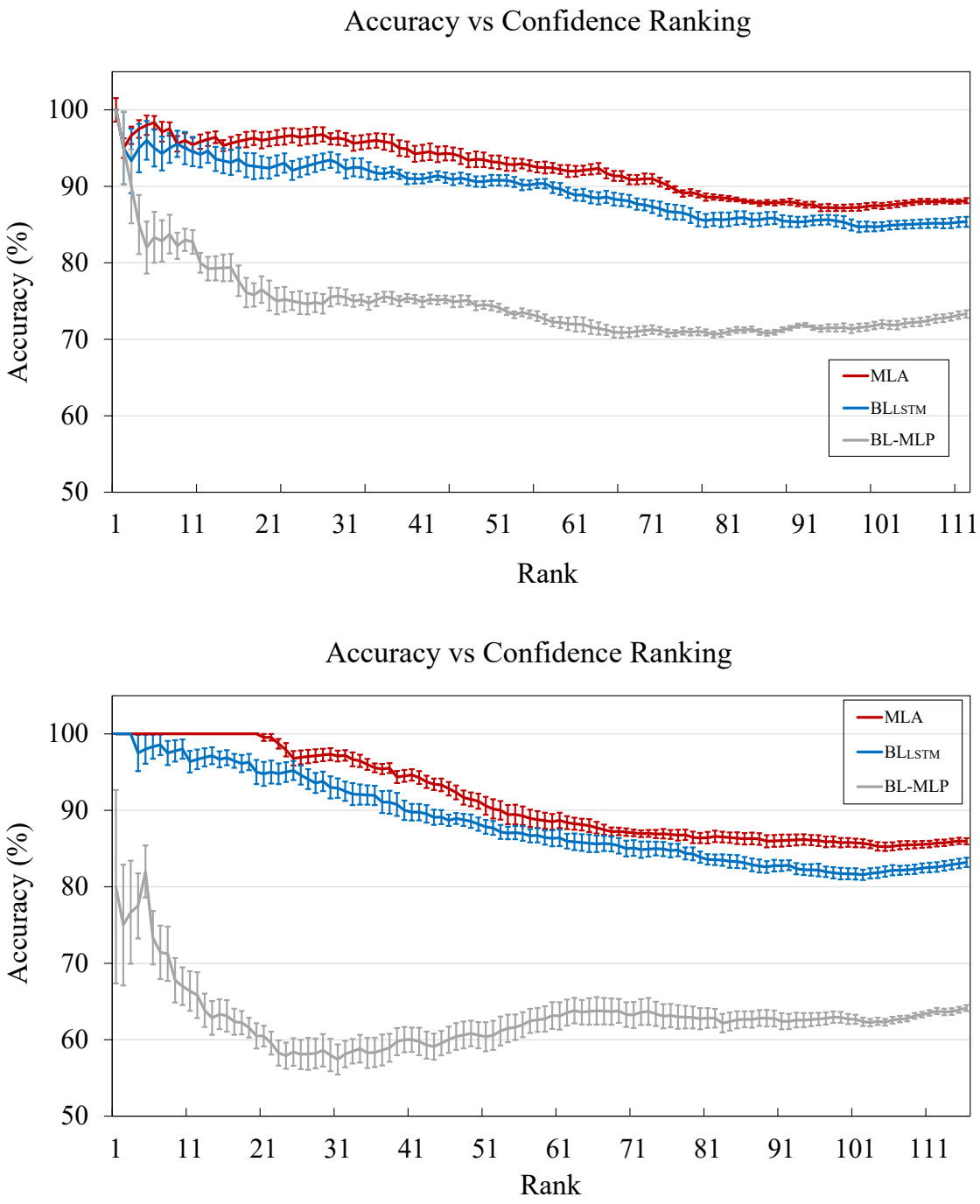

Figure 6.2: The figure shows the accuracy obtained when considering only the speakers showing the r highest confidence scores. The vertical bars correspond to the standard error of the mean observed across R repetitions. The upper plot corresponds to the results of the Read Task while the lower plot corresponds to the results of the Interview Task.

resources, as the response of the system can be accepted when s is high enough, while requesting the attention of a doctor when s is too small. As a consequence, the workload for the doctors can be reduced while keeping the accuracy of the diagnosis approach high enough. In this respect, the application of MLA improves confidence in the results given by the automatic depression

detector, more effectively alleviating the workload for doctors.

The Read Task yields a more reliable confidence score and exhibits less variance compared to the Interview Task when considering all speakers. This outcome may be attributed to the fact that the Read Task requires participants to read predefined text, leading to less variability in the speech samples. However, the use of the Interview Task allows a larger subset of participants corresponding to $s > 99\%$. This result seems to suggest the possibility of combining Read and Interview Tasks in automatic depression detection, i.e., employing the Interview Task for participants with sufficiently large s values and utilizing the Read Task to detect the remaining participants. This combined approach could be a future direction in depression detection and may further enhance the overall performance of the automatic depression detection system.

6.4.3 Number of Frames

A crucial research question addressed in this study concerns the efficiency of the proposed approach in detecting depression, specifically within the context of tasks such as the Interview Task. Our objective is to determine whether it is necessary to analyze the entirety of the recording or if examining only part of it would yield similar results. The main motivation behind this is to investigate the feasibility of reducing the length of the test and, as a result, reducing the effort of the patients that take it. This is of particular importance for individuals with severe depression who might not be at ease while undergoing a test that takes a long time.

Figure 6.3 shows how the accuracy changes as a function of the number of frames used to perform the classification. Upon the comparison between the curves, it becomes evident that the MLA approach substantially outperforms the other approaches, as indicated by a statistically significant difference ($p < 0.001$, according to a two-tailed t-test) for every number of frames in both types of speech. Given that increasing the number of frames means to increase the amount of speech time used to classify a speaker, this means that MLA can reach a predefined accuracy earlier than the baselines. The relationship between the number of frames, n, and the recording duration (seconds) can be expressed as:

$$\iota = (n-1) \times 0.65 + 1.3, \tag{6.4}$$

where ι is the number of recording, refer to Section 3.3 for further details about frame size and step.

A comprehensive analysis of the plots suggests that the results do not exhibit significant fluctuations after the initial 23 frames, corresponding to the first 15.6 seconds of the Interview Task. Given that the average length of the Interview Task is 229.8 seconds (see Section 3.3), this leads to a substantial 93.2% reduction in the duration of data required to make a prediction. Conversely, the Read Task seems to demand a greater number of frames (41 frames) before the results reach a stable point, corresponding to 27.3 seconds (compared to an average Read

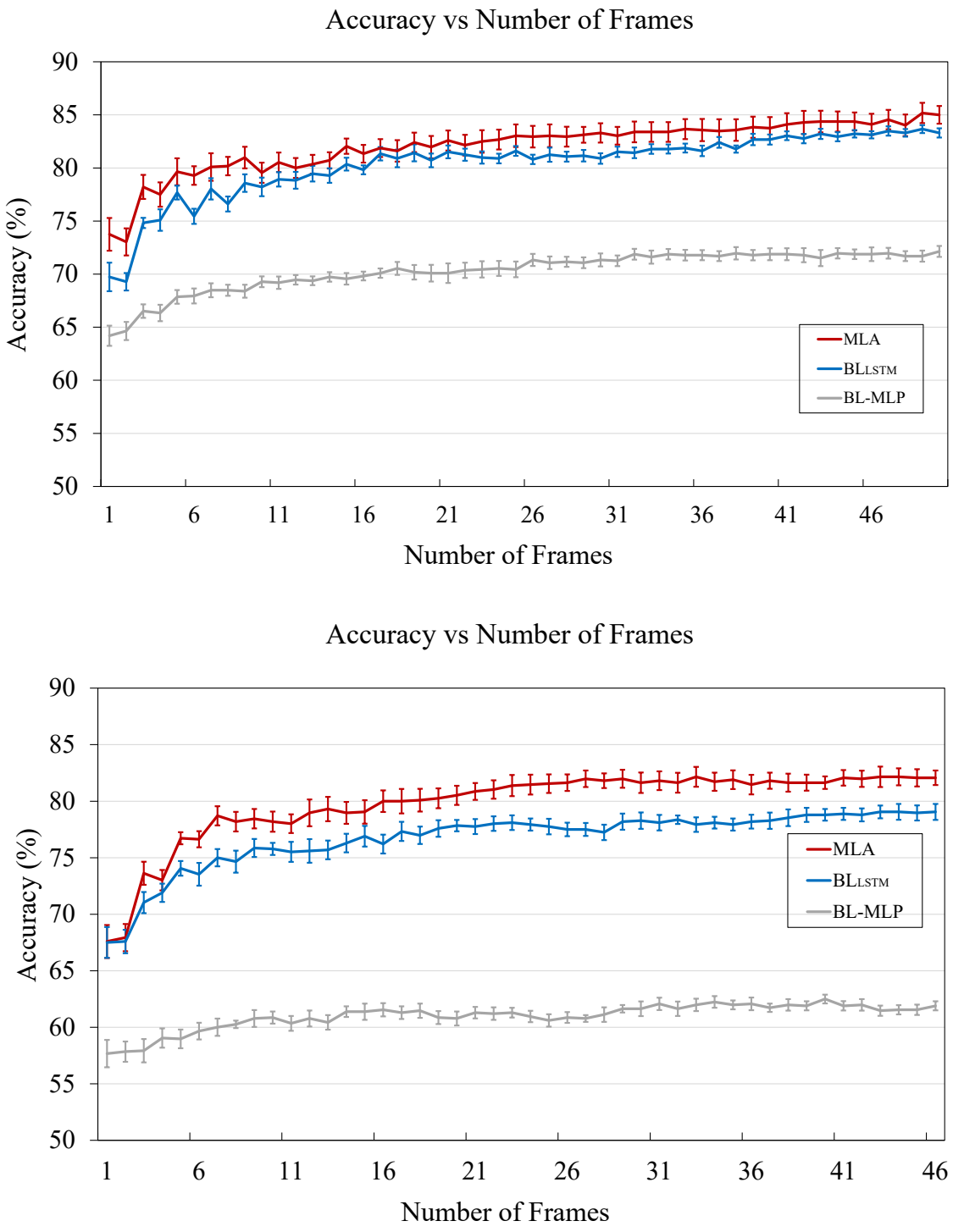

Figure 6.3: The figure shows the relationship between accuracy and the number of frames (50 and 46 are the maximum of frames that every participant has for Read Task and Interview Task, respectively) used for depression detection. The vertical bars correspond to the standard error of the mean observed across R repetitions. The upper plot corresponds to the results of the Read Task while the lower plot corresponds to the results of the Interview Task.

Task length of 50.3 seconds). This observation results in a 49.8% reduction in the duration of

data required to make a prediction. In conclusion, our findings demonstrate that focusing on the initial few seconds of read and spontaneous speech provides sufficient data for achieving performance comparable to that obtained when utilizing the entire available material.

6.5 Conclusions

In this chapter, we answered the research question proposed in Section 6.1. We present a comprehensive analysis demonstrating that the performance of a depression detector can be significantly enhanced through the application of Multi-Local Attention. This attention mechanism is specifically designed to emphasize input data anticipated to convey task-relevant information, thereby improving the accuracy and reliability of the detection process. Moreover, our findings reveal two additional and noteworthy improvements in the performance of the detector that hold significant implications for its practical application in the field of mental health care.

Firstly, we observe that the confidence measure accompanying detection outcomes becomes more effective in accurately identifying classified speakers. This improvement is of particular importance because it makes it possible to identify participants for which the actual condition, whether it be depression or lack of it, is evident enough to be recognized automatically. By effectively identifying these individuals, doctors can allocate their time and resources more efficiently, concentrating on more difficult and ambiguous cases.

Secondly, our results show that the detector achieves its best accuracy by using a smaller amount of speech data, which is important considering that depressed individuals often struggle to speak for a long time. The experiments show that approximately 15 to 25 seconds of two different types of speech are sufficient for detecting depression as effectively as when utilizing all available data. Compared to the initial approaches, the proposed approach reduces the duration of data required to make a prediction but makes a relatively accurate detection, which would be a benefit if deployed in real-world systems. Moreover, this approach ensures that the most vulnerable individuals, such as severely depressed individuals who find it difficult to speak for lengthy periods, are not subjected to time-consuming and potentially distressing testing procedures.

In addition, the use of Read and Interview Tasks exhibits distinct patterns in the aforementioned aspects. This observation seemingly confirms that read and spontaneous speech convey different types of information, a conclusion that is consistent with previous studies in the field (Mitra & Shriberg, 2015; Kiss & Vicsi, 2017a). As a result, our focus in the subsequent chapter will be on developing improved methodologies for effectively combining Read and Interview Tasks, thereby capturing such information in different tasks for further enhancing the accuracy and reliability of depression detection.

Another positive effect of the Multi-Local Attention is that the fraction of correctly classified frames tends to increase (hence the higher confidence scores for the correctly classified speakers). Such an observation is consistent with the thin slices theory (the tendency of people

to manifest their inner state through short behavioral displays) (Ambady & Rosenthal, 1992) and it is the probable explanation behind the effectiveness of the majority vote. Given that such an aggregation approach is basic, we will focus on the attempt to develop better methodologies to combine the classification outcomes obtained at the level of individual frames in the next chapter.

Chapter 7

Cross-Data Multilevel Attention

7.1 Introduction

In the previous chapters, we compared the effectiveness of markers and models applied to Read and Interview Tasks in the context of depression detection. The experiments show the differences in effectiveness when using different types of speech, i.e., read and spontaneous speech. This is mainly due to the way that speech is produced, in which read speech limits the content of speaking representing more difference in acoustics while spontaneous speech can express the emotional state of the speaker. So far, the experiments were based on comparisons and no attempt was made to combine two types of speech. This leads to the following questions: *can different types of speech convey different information for depression detection, and if so, can the combination of this different information improve the performance of depression recognition?*

In this chapter, we proposed a new approach to analyse both read and spontaneous speech in an individual framework using several attention mechanisms used in the last chapter. These latter operated at multiple levels in terms of time-scale (from short frames of 1.3 seconds to entire recordings lasting up to several minutes) and data representation (from sequences of feature vectors extracted from the speech signal to sequences of embeddings resulting from Deep Networks). The goal of the attention mechanisms is to leverage depression-relevant information available in both read and spontaneous speech, whether it is specific to each type of data or common to both.

The literature proposes three main ways to collect speech data for depression detection, namely asking people to read a text, posing a series of simple questions or asking to talk as much as possible about a predefined topic (e.g., the description of a picture). The first approach leads to read speech, while the other two lead to spontaneous speech. The literature has extensively analyzed advantages and disadvantages of both types of data (see, e.g., (Kiss & Vicsi, 2017a)). Overall, the results seem to suggest that spontaneous speech allows one to obtain better performance (Alghowinem, Goecke, Wagner, Epps, Breakspear, & Parker, 2013; Mitra & Shriberg, 2015). However, the differences are small (accuracy differences range between 1%

and 3%) (Kiss & Vicsi, 2017a). As a consequence, several works suggest that the best solution is to use both types of data (Liu, Li, et al., 2017; Mitra et al., 2015; Long et al., 2017). This work follows such a methodology and proposes a new depression detection approach capable to benefit from both read and spontaneous speech samples of the same speaker.

The proposed approach, referred to as *Cross-Data Multilevel Attention* (CDMA), relies on the key-assumption that both read and spontaneous speech carry depression-relevant information. However, while part of such information is specific of read or spontaneous speech, the other part is common to both. For this reason, the CDMA includes multiple attention mechanisms that are both type-specific, meaning that they aim at emphasizing information available only in read or spontaneous data, and type common, meaning that they try to emphasize information available in both types of speech. The experiments show that removing one or more of the attention mechanisms of the CDMA reduces the performance. Such a result seems to confirm that the key-assumption above is correct and that the CDMA is effective at leveraging all information available in the data.

The experiments were performed over the Androids Corpus including 52 individuals diagnosed with depression by professional psychiatrists and 58 individuals who had never experienced mental health issues. During the experiments, each participant was asked to read a simple text and to answer a few questions. In this way, both read and spontaneous speech samples were available for each person. The 8 participants for which only one type of data was available were excluded (see Chapter 3 for more data details). The experiments show that the CDMA can achieve F1-Scores of up to 92.5%, the best result obtained so far over the same data.

Overall, to the best of our knowledge, the main contributions of the work in this chapter are as follows:

- A new approach, the CDMA, that was developed in a depression detection experiment, but can be used to analyze all pairs of data items that share a common label in a classification problem and can be represented as sequences of feature vectors;
- Experimental evidence that, in depression detection, read and spontaneous speech carry both specific and common information and that these are sufficiently diverse to improve the performance of a detection approach when combined.

The rest of this chapter is organized as follows: Section 7.2 provides information about the data used for the experiments, Section 7.3 illustrates the proposed approach, Section 7.4 describes experiments and results, and the final Section 7.5 draws some conclusions.

7.2 The Data

The experiments involved a subset of the Androids Corpus including 110 participants, of which 52 participants are *control* while 58 participants are *depressed*. The only reason for participant

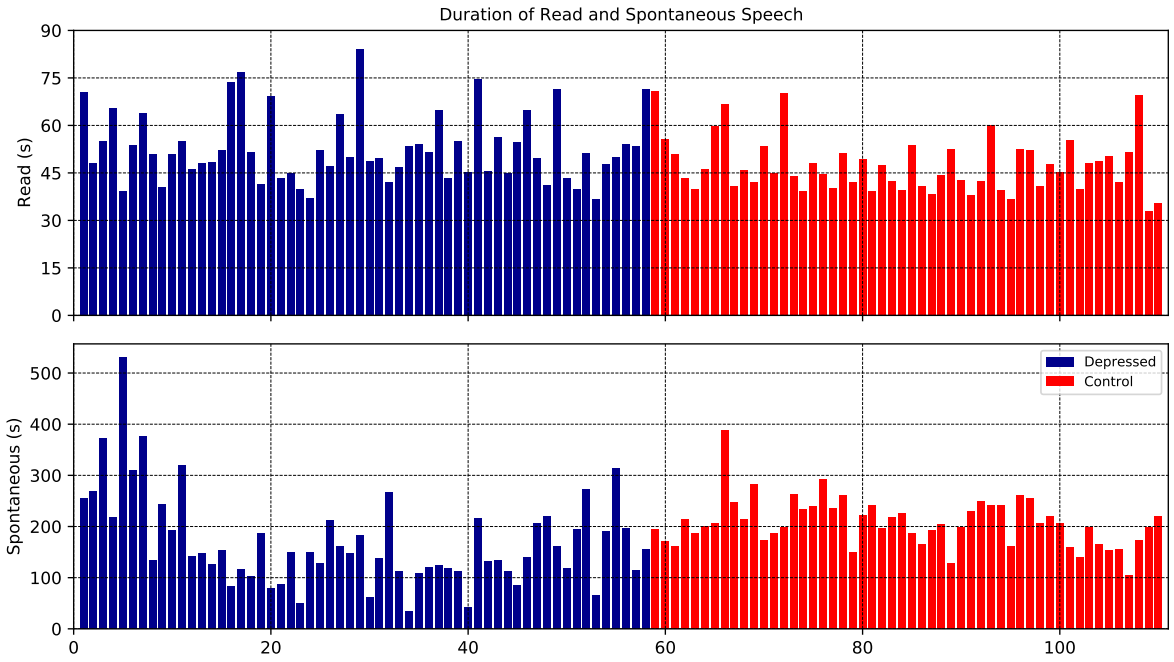

Figure 7.1: Distribution of recording length across the participants with both types of data. The upper chart shows the length of read speech, and the lower shows the same information for spontaneous speech (after removing the turns of the interviewer).

selection from the Androids Corpus is that the Read and Interview Tasks should be available. Therefore, a total of 8 participants are excluded due to the lack of either the Read Task or Interview Task.

Table 7.1 shows the distribution of age, gender and education level across Control and Depressed participants in the subset. According to a two-tailed t-test, there is no statistically significant age difference between the two groups. The same applies to gender and education level, according to a χ^2 test. This ensures that differences between the two groups of participants depend on depression and not on other factors that might affect speech. In addition, there is no statistically significant difference between the subset used in this chapter and the Androids Corpus in terms of age (according to a two-tailed t-test), gender and education level (according to χ^2 tests). This means the results obtained from this subset can be generalized to the whole Androids Corpus.

In the Reading task, the total duration of the recordings is 1 hour, 31 minutes and 43 seconds (the upper chart of Figure 7.1 shows the duration per participant). The overall average and standard deviation are 50.3 ± 10.4 seconds, while average and standard deviation for depressed and control participants are 52.9 ± 10.9 and 47.3 ± 8.9 seconds, respectively. According to a two-tailed t test, such a difference is statistically significant ($p < 0.01$). In the Interview recordings, the total duration of the turns is 5 hours, 44 minutes and 58 seconds, and the lower chart of Figure 7.1 shows the duration per participant (the overall average and standard deviation

Table 7.1: The table provides demographic information about the participants in terms of age, gender and education level. The expressions *Low* and *High* refer to this latter. Low means up to 8 years of study, while High means at least 13 years of study. The sum over the education level columns is only 109 because 1 participants did not disclose information about their studies.

	Age	Male	Female	Low	High
Control	47.3 ± 12.7	11	41	19	33
Depressed	47.4 ± 11.9	20	38	25	32
Total	47.3 ± 12.2	31	79	44	65

are 188.2 ± 76.2 seconds). The average and standard deviations for depressed and control participants are 170.5 ± 91.8 and 207.9 ± 47.3 seconds. The difference is statistically significant between depressed and control participants according to a two-tailed t test ($p < 0.01$). These statistical differences and the distribution of recording length in the subset are in line with the description of the whole corpus in Section 3.2.

7.3 The Approach

The proposed approach includes four main steps, namely *feature extraction*, *segmentation*, *detection* and *aggregation* (see Figure 7.2 for the overall scheme).

7.3.1 Feature Extraction

The feature extraction step is the same as the baselines described in Section 3.3, including the feature set, the size of analysis window and tools for extracting features.

Figure 7.2 shows that the output of the feature extraction step are a sequence of feature vectors $X = \{\vec{x}_1, \ldots, \vec{x}_T\}$ for read speech (T is the total number of feature vectors) and a sequence of feature vectors $Z = \{\vec{z}_1, \ldots, \vec{z}_Q\}$ for spontaneous speech (Q is the total number of feature vectors). In general, $T \neq Q$ because the amount of read and spontaneous speech is different for the same speaker.

7.3.2 Segmentation

After the feature extraction step, each recording corresponds to a sequence of feature vectors (see the end of previous section). Given the same problem as the baselines described in Section 3.3.4 that minimum recording length is in the order of tens of seconds, all sequences include thousands of vectors, and this makes it impossible to feed them to a model. Therefore, the sequences X and Z are segmented into frames following the same way for baselines in Section 3.3.4. In the case of read speech (Read Task), this means to convert X into a sequence $I = \{I_k\}$ ($k \in [1, N]$), where

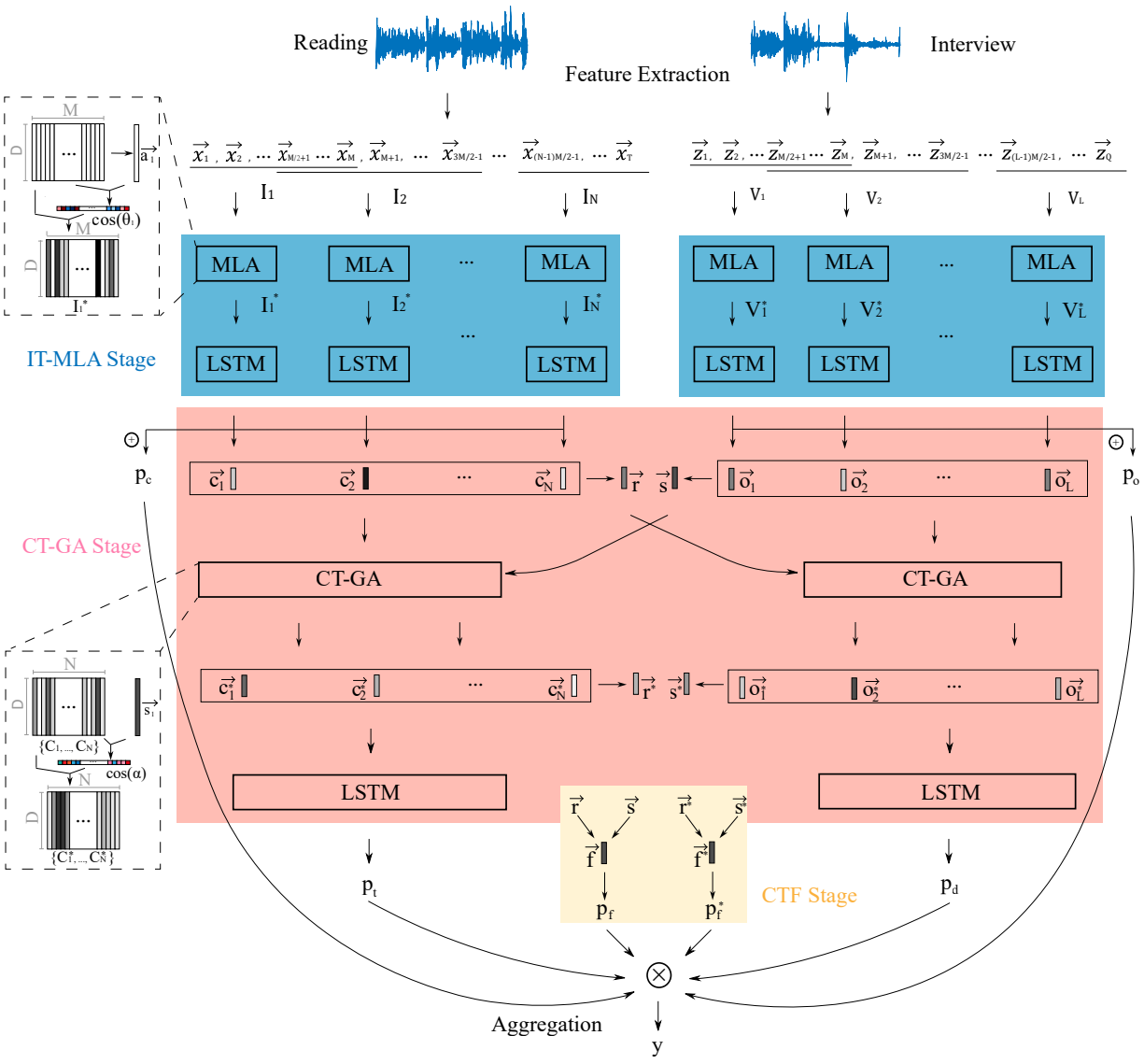

Figure 7.2: The figure shows the main stages of the proposed approach. The symbol ⊗ corresponds to the aggregation of the outcomes produced by the different stages of the approach (see the text for the meaning of symbols).

every frame I_k includes $M = 128$ vectors (see Section 3.3.4). The frames also start at regular steps of length $M/2$, meaning that they overlap by half of their vectors (see Figure 7.2).

In the case of spontaneous speech (Interview Task), the segmentation into frames was applied individually to each turn of the participants (the turns of the interviewers were manually eliminated from the data). In this way, no frame includes vectors extracted from different turns. In technical terms, this means that the initial sequence of vectors $Z = \{\vec{z}_1, \ldots, \vec{z}_Q\}$ extracted from an Interview recording was first segmented into turns and then each of these was segmented individually into frames according to the same procedure used for the read speech. As a result, spontaneous speech recordings are converted into sequences $V = \{V_m\}$ ($m \in [1, L]$) of frames.

Overall, the final outcomes of the segmentation step are a sequence of frames I for read speech and a sequence of frames V for spontaneous speech (see Figure 7.2).

7.3.3 Detection: Intra-Type Multi-Local Attention

After the segmentation step, read and spontaneous speech are represented as sequences of frames. The first stage of the detection step, referred to as *Intra-Type Multi Local Attention* (IT-MLA) focuses on each type of speech individually (hence the expression *Intra-Type*) and tries to emphasize local depression-relevant information available in each frame (hence the expression *Multi-Local Attention*). The IT-MLA is the same as MLA in Chapter 6 Section 6.3. In this chapter, we use IT-MLA to replace MLA in order to help readers understand this attention mechanism is based on intra-type, different from the other attention mechanism that is based on cross-type proposed in this chapter (see below). The calculation of IT-MLA is the same as MLA in Section 6.3.2 that uses the same equation between the average feature vector $\vec{a}_k$ from every frame k and the vectors $\vec{x}_i$ of the frame k (see Equation 6.1 and Equation 6.2, the equation is written for the vectors $\vec{x}_i$ extracted from read speech, but it can be written in the same way for the vectors $\vec{z}_i$ extracted from spontaneous speech).

After the transform, the frames can be given as input to Long Short-Term Memory Networks (LSTMs) (Hochreiter & Schmidhuber, 1997) to be assigned either to class *depression* or to class *control*. Given that there are multiple frames, there are multiple classification outcomes and it is possible to estimate the probability of a speaker being depressed as follows:

$$p(d) = \frac{n(d)}{N}, \tag{7.1}$$

where $n(d)$ is the number of frames assigned to class *depression* and N is the total number of frames. This leads to probabilities p_c and p_o of a speaker being depressed based on read and spontaneous speech, respectively (see Figure 7.2). When $p_c > 0.5$, a speaker is classified as depressed for read speech (the same applies to p_o for spontaneous speech).

7.3.4 Detection: Cross-Type Global Attention

The second stage of the detection step, referred to as *Cross-Type Global Attention* (CT-GA), aims at jointly using information available in both types of speech (hence the expression *Cross-Type*) and at emphasizing information available at the level of a whole recording (hence the expression *Global Attention*).

Whenever an LSTM used for IT-MLA classifies a frame (see previous section), it produces a sequence of hidden states. It is then possible to calculate the average of the hidden states so that every frame leads to an individual vector $\vec{c}_k$ for read speech and $\vec{o}_k$ for spontaneous speech. Given that there are multiple frames, the LSTMs of the IT-MLA produce two sequences, namely

$C = \{\vec{c}_1, \ldots, \vec{c}_N\}$ for read speech and $O = \{\vec{o}_1, \ldots, \vec{o}_L\}$ for spontaneous speech (see Figure 7.2).

The CT-GA stage of the detection transforms the two sequences above according to the following equation (the expression is written for the $\vec{c}_k$, but it can be used in the same way for the $\vec{o}_k$):

$$\vec{c}^*_k = \vec{c}_k + \frac{exp(\cos \alpha_k)}{\sum_{j=1}^{N} exp(\cos \alpha_j)} \cdot \vec{c}_k, \tag{7.2}$$

$$\cos \alpha_k = \vec{s}\vec{c}_k / ||\vec{s}||||\vec{c}_k||, \tag{7.3}$$

where $\vec{s}$ is the average of O. In the case of spontaneous speech, it is necessary to use the average $\vec{r}$ of sequence C. In sequence O, the transformation above emphasizes the parts that are more similar to the average of C and the other way around. This means that the transform emphasizes in read speech what is similar to spontaneous speech and vice versa. The expected consequence is that the transformed sequences convey information common to both types of speech.

The transformed sequences C^* and O^* are fed to two LSTMs (see Figure 7.2) that generate a new hidden state for each input vector $\vec{c}^*_k$ and $\vec{o}^*_k$, respectively. This makes it possible to calculate the average of the hidden states for each of the two LSTMs. Such averages are fed to two softmax layers that give as output the probabilities p_t and p_d of the speakers being depressed, respectively (see Figure 7.2).

7.3.5 Detection: Cross-Type Fusion

After the CT-GA step, there are four sequences, namely C, O, C^* and O^*, with their respective averages $\vec{r}$, $\vec{s}$, $\vec{r}^*$ and $\vec{s}^*$ (see Figure 7.2). This makes it possible to obtain two vectors as follows:

$$\vec{f} = \vec{r} + \vec{s}, \tag{7.4}$$

$$\vec{f}^* = \vec{r}^* + \vec{s}^* \tag{7.5}$$

where vectors $\vec{f}$ and $\vec{f}^*$ account for information that was available in the intermediate steps of the process because $\vec{f}$ was obtained immediately after IT-MLA, and $\vec{f}^*$ was obtained during CT-GA. To ensure that such information does not get lost, the two vectors are fed to two softmax layers that output two probabilities p_{f1} and p_{f2} of the speakers being depressed. The rationale behind the step is that many approaches detect depression above chance by representing speech recordings through the average of feature vectors extracted at regular time steps from the signal (see, e.g., (Tao et al., 2020; Kiss & Vicsi, 2017a)). This suggests that the averages $\vec{r}$, $\vec{s}$, $\vec{r}^*$ and $\vec{s}^*$, convey information about the condition of the speaker.

Given that $\vec{f}$ and $\vec{f}^*$ account for averages extracted from different types of speech, the step is referred to as *Cross-Type Fusion* (CTF).

7.3.6 Aggregation

The six probabilities estimated during the process are aggregated in terms of the average:

$$\hat{p} = \frac{1}{|\mathscr{B}|} \sum_{v \in \mathscr{B}} p_v, \tag{7.6}$$

where $\mathscr{B} = \{c, o, f_1, f_2, t, d\}$. Whenever $\hat{p} > 0.5$, a speaker is classified as depressed. In addition, the probabilities are used to jointly train LSTMs and softmax layers (see previous sections) according to the following loss function:

$$\mathscr{L} = -\frac{1}{K} \sum_{v \in \mathscr{B}} [y \log(p_v) + (1-y) \log(1-p_v)], \tag{7.7}$$

where K is the number of training samples.

7.4 Experiments and results

All experiments were performed according to a k-fold protocol ($k = 3$). This is in line with the experimental design of baselines in Section 3.3, following the rule of *person independence* that the same person is never represented in both training and test set. All the models were trained on the same devices as baselines (a single Tesla T4 GPU with 16 GB memory). The number of hidden neurons in the LSTMs was set to 32, the learning rate to 10^{-3} and the number of training epochs to 300. The models were trained with the initializer of PyTorch using RMSProp as an optimizer (Tieleman et al., 2012). Because of the random initialization of the weights, the experiments were repeated $R = 10$ times and, therefore, the performance metrics are reported as averages and standard deviations across the R repetitions.

Table 7.2 shows the results obtained in this work. The first line is the performance of a random baseline approach that assigns a sample to the class c with a probability corresponding to its prior $p(c)$, in line with the Equation 3.7 in Chapter 3. Precision, Recall and F1-Score of the same random approach correspond to the prior of the positive class (*depression* in this chapter). The remaining lines of the table account for approaches that include progressively more stages of the CDMA (see below).

According to a two-tailed t-test with FDR correction ($p < 0.001$), the performance is always better than the random baseline to a statistically significant extent. This means that all stages of the CDMA actually learn from the data to discriminate between depressed and non-depressed speakers. The full CDMA approach (lowest line of Table 7.2) outperforms, to a statistically significant extent, all partial approaches that include only some of its stages. In this respect, all stages of the CDMA (IT-MLA, CT-GA and CTF) appear to contribute with diverse and alternative information to the final outcome of the process.

Table 7.2: The table provides recognition results in terms of Accuracy, Precision, Recall and F1 Score (see the text for the meaning of the acronyms).

Approaches	Type		Stages						Acc.(%)	Pre.(%)	Rec.(%)	F1.(%)
	Read	Spont.	MLA	LSTM1	GA		LSTM2	CTF				
					self	cross		f1 f2				
Random									50.1±0.0	52.7±0.0	52.7±0.0	52.7±0.0
Part I												
BA1 (Read)	√			√					85.5±1.2	84.1±2.6	86.8±2.3	84.7±1.3
BA1 (Spont.)		√		√					83.1±0.8	80.4±2.1	87.8±2.7	83.3±0.7
IT-MLA (Read)	√		√	√					89.2±0.7	88.2±1.8	90.1±1.6	89.0±0.6
IT-MLA (Spont.)		√	√	√					86.8±1.1	83.9±2.3	91.9±1.5	87.4±0.8
Part II												
BA2 (Reading)	√		√	√			√		89.9±1.5	88.7±2.1	91.5±2.8	89.9±1.5
BA2 (Interview)		√	√	√			√		87.3±1.1	85.8±2.4	89.4±3.3	87.3±1.1
BA3 (Read)	√		√	√	√		√		90.1±1.6	88.7±3.2	92.4±2.6	90.2±1.2
BA3 (Spont.)		√	√	√	√		√		89.4±0.9	87.1±1.8	93.0±2.5	89.7±0.8
CT-GA (Read)	√		√	√		√	√		90.4±1.2	88.4±2.0	93.6±2.3	90.7±1.0
CT-GA (Spont.)		√	√	√		√	√		89.8±0.8	88.8±1.6	91.4±1.9	89.8±0.7
Part III												
BA4	√	√	√	√		√	√	√	90.2±0.8	88.9±1.7	91.5±1.5	90.1±0.8
BA5	√	√	√	√		√	√	√	90.7±0.6	89.3±2.0	92.5±2.8	90.7±0.6
CDMA	√	√	√	√		√	√	√ √	**92.7±0.8**	**91.8±1.4**	**93.5±2.1**	**92.5±0.8**

7.4.1 Performance Analysis

Figure 7.2 shows that the CDMA can be thought of as a sequence of multiple stages. The goal of this section is to show how the performance changes along such stages and, correspondingly, Table 7.2 provides Accuracy, Precision, Recall and F1-Score when taking into account increasingly more of them. This provides an indication of how each of the stages contributes to the overall performance of the CDMA, reported in the lowest line of the table.

IT-MLA

Part I of Table 7.2 shows a comparison between the IT-MLA (see Section 7.3.3) and a baseline approach, referred to as BA1, that feeds the sequences of frames $I = \{I_k\}$ or $V = \{V_m\}$ to an LSTM (see Figure 7.2). The only difference between BA1 and IT-MLA is that this latter applies the MLA before feeding the data to an LSTM. Both IT-MLA and the baseline were trained and tested separately over read and spontaneous speech. The MLA increases the accuracy by 3.7% for both types of speech, thus reducing the error rate by 25.5% for read speech and by 21.9% for spontaneous speech. Such improvements are statistically significant with $p < 0.001$ according

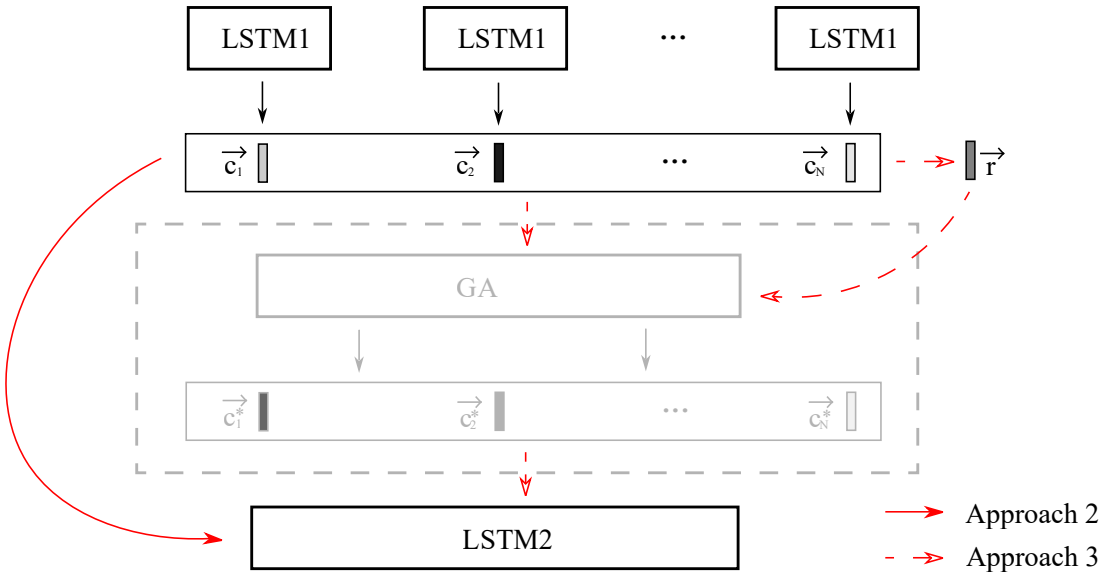

Figure 7.3: The plot shows BS2 and BS3. In BS2, the sequence C is fed to the LSTMs skipping the global attention stage to generate a posterior for depression detection. In BS3, the sequence C is applied with global attention to generate sequence C^*, then fed C^* to LSTMs to generate a posterior for depression detection. These steps also works for the sequence O.

to two-tailed t-tests (after applying the FDR correction). This suggests that the MLA effectively emphasizes depression-relevant information in each of the data types.

CT-GA

Section 7.3.4 shows that the CT-GA stage transforms sequences C and O and then feeds the transformed sequences C^* and O^* to two LSTMs (represented as LSTM2 in Figure 7.2). This leads to the estimates of two posteriors, p_t and p_d in Figure 7.2, that can be used to classify a speaker. Part II of Table 7.2 shows the performance achieved when using the two posteriors individually to perform the classification (the corresponding approaches are referred to as CT-GA over read and spontaneous speech, respectively).

The first key-point of the CT-GA is that sequences C and O are transformed according to global attention. For this reason, it is possible to compare the performance obtained by the CT-GA with the baseline (referred to as BA2) that eliminates global attention, meaning sequences C (or O) are fed to the LSTMs are not transformed (skipping GA in the BA2). This comparison aims at examining the effectiveness of global attention.

The second key-point of the CT-GA is that sequence C is transformed using the average of sequence O and vice versa (see Equation 7.3). This aims at emphasizing information common to both types of speech while still processing individually each of them. For this reason, it is possible to compare the performance obtained by the CT-GA with an alternative baseline

approach, referred to as BA3, that transforms sequences C and O as follows:

$$\vec{c}_k^* = \vec{c}_k + \frac{exp(\cos\beta_k)}{\sum_{i=1}^{N} exp(\cos\beta_i)} \cdot \vec{c}_k, \tag{7.8}$$

$$\vec{o}_k^* = \vec{o}_k + \frac{exp(\cos\gamma_k)}{\sum_{i=1}^{N} exp(\cos\gamma_i)} \cdot \vec{o}_k, \tag{7.9}$$

where β_k is the angle between c_k and the average of sequence C, while γ_k is the angle between o_k and the average of sequence O. This is the same transform as the one used in the CT-GA stage (see Section 7.3.4). However, while the CT-GA uses the average of O to transform C and vice versa, the transform above use the average of C to transform C and the average of O to transform O. In other words, there is no attempt to use information common to the two types of speech. The transform is Intra-Type rather than Cross-Type.

Overall, BA3 and CT-GA perform at the same level according to a two-tailed t-test (none of them outperforms the others to a statistically significant extent). However, it is important to observe that, compared to BA2 in Table 7.2, there is a statistically significant improvement for spontaneous speech according to a two-tailed t-test ($p < 0.001$ with FDR correction). This seems to suggest that the GA allows one to bridge the performance gap between read and spontaneous speech observed earlier. One possible explanation is that GA emphasizes information that spans an entire recording. Therefore, it is more likely to emphasize depression-relevant differences in spontaneous speech (where every speaker says a different thing) than in read speech (where every speaker reads the same text).

CTF

Section 7.3.5 shows that the Cross-Type Fusion is performed by feeding vectors $\vec{f}$ and $\vec{f}^*$ to two softmax layers that output the posteriors p_{f1} and p_{f2} of class depression (see Figure 7.2). These are summed to the posteriors resulting from IT-MLA (p_c and p_o) and from CT-GA (p_t and p_d), thus resulting in the full CDMA approach.

Part III of Table 7.2 shows the comparison between the CDMA and two alternative baseline approaches, BA4 and BA5. These correspond to the full CDMA, but BS3 does not take into account p_{f2}, while BS4 does not take into account p_{f1}. The key-result is that the CDMA improves to a statistically significant extent over all approaches of Parts I-III in Table 7.2, while baselines BA4 and BA5 improve only over the approaches of Part I of the table. One possible explanation is that the two fusion approaches contribute with diverse information, meaning that what is available in $\vec{f}$ is not available in $\vec{f}^*$ and vice versa. For this reason, using only one of the two is not sufficient and it is necessary to use them both. In any case, adding a stage that takes information coming from both types of speech leads to the best performance, thus confirming the key-assumption underlying this work that both read and spontaneous speech have to be taken into account.

Table 7.3: Previous results over same data. R and S stand for Read and Spontaneous, respectively (S+Text means that both spontaneous speech and its transcription were used).

Methods	Task	Acc.(%)	Pre.(%)	Rec.(%)	F1(%)
Scibelli et al. (2018)	R	77.0	74.0	80.0	77.0
Tao et al. (2020)	R	84.5	84.5	84.6	84.5
Tao, Ge, et al. (2023)	R	88.0	87.7	89.0	88.0
Aloshban et al. (2020)	S+Text	83.5	95.0	70.3	80.5
Aloshban et al. (2021)	S+Text	83.0	95.2	69.0	80.0
Aloshban et al. (2022)	S+Text	84.7	**95.4**	72.4	82.3
Alsarrani et al. (2022)	S	67.6	71.7	72.2	72.3
This work	R+S	**92.7**	91.8	**93.5**	**92.5**

7.4.2 Comparison with Other Works

Table 7.3 shows the comparison between the CDMA and previous approaches tested over the same data. The approaches in the table do not make use of the same experimental protocol used in this work and they all use different subsets of the data (e.g., only read or only spontaneous speech). Therefore, the comparisons are not fully rigorous and should be considered only as an indication.

According to a two-tailed t-test with *False Discovery Rate* (FDR) correction, the CDMA outperforms the other methods to a statistically significant extent ($p < 0.001$ in all cases). Three studies (Aloshban et al., 2020, 2021, 2022) report Precision higher than CDMA, but all of them have a Recall roughly 20 points lower and they were carried out using only a small subset of the data (59 speakers in total). Furthermore, the comparison in terms of F1-Score suggests that the CDMA tends to achieve a better balance between Precision and Recall. Finally, the results suggest that taking into account both read and spontaneous speech for the same speaker is of benefit (all other results were obtained using only one type of speech).

7.5 Conclusions

This chapter presented an approach, the *Cross-Data Multilevel Attention*, aimed at detecting depression through the analysis of read and spontaneous speech. The performance gap between RT and IT may be attributed to the different nature of these two speech types, each containing different sets of information. Unlike previous approaches that analyze both read and spontaneous speech (see, e.g., (Alghowinem, Goecke, Wagner, Epps, Breakspear, & Parker, 2013; Kiss & Vicsi, 2017a; Liu, Li, et al., 2017; Long et al., 2017; Mitra & Shriberg, 2015; Mitra et al., 2015)), the main novelty of the CDMA is that, through the use of attention mechanisms, it tries to explicitly capture all depression-relevant information available in the data, whether it is

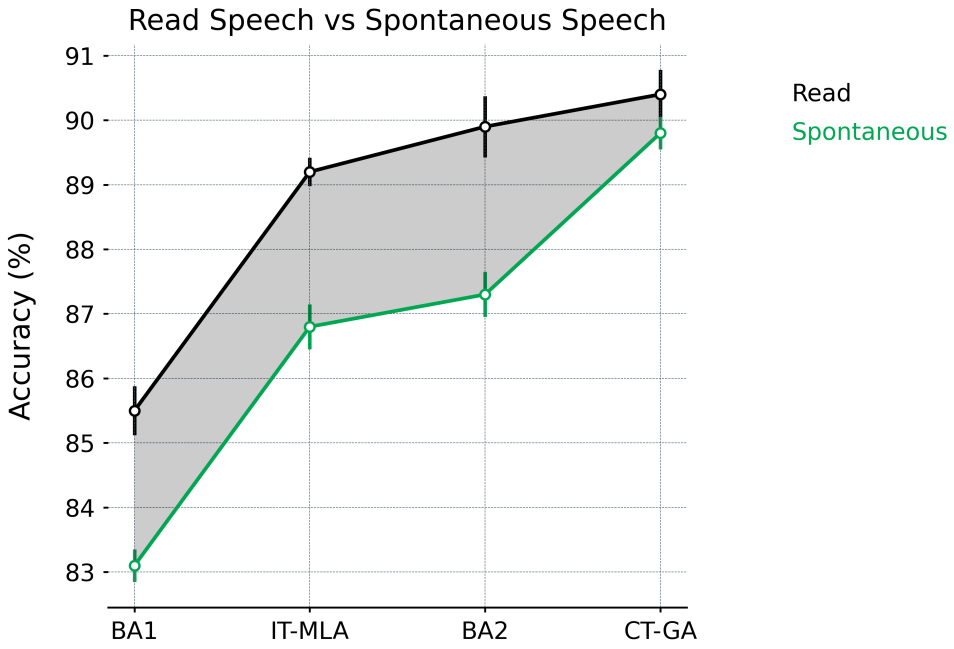

Figure 7.4: This figure shows the comparison between Read and Spontaneous speech in terms of accuracy (the vertical bars correspond to the standard error of the mean performance across the R repetitions of the experiments).

specific of each type of speech or common to both of them. The results show that the inclusion of progressively more attention mechanisms consistently increases the performance over both types of speech. Table 7.2 shows that, after applying *Multi Local Attention* and *Global Attention*, the F1-Score increases by 6.0% for read speech and by 6.5% for spontaneous speech. Furthermore, the results show that using jointly read and spontaneous speech leads to a statistically significant improvement over the best results obtained using only one type of data. In fact, the CDMA reaches an F1-Score of 92.5%, while the best F1-Scores obtained using individually read or spontaneous speech are 90.7% and 90.2%, respectively (see Table 7.2).

In line with previous works (see, e.g., (Alghowinem, Goecke, Wagner, Epps, Breakspear, & Parker, 2013; Kiss & Vicsi, 2017a)), there is a performance difference of roughly 3.0% between read and spontaneous speech. However, before the application of the GA, the best performance was obtained over read speech (see Figure 7.4), unlike all previous works (see, e.g., (Liu, Li, et al., 2017; Mitra & Shriberg, 2015)). One possible explanation is that the way the MLA works tends to benefit more from read speech. In fact, the key-assumption underlying the MLA is that the average of the feature vectors extracted from a sample conveys depression-relevant information. In the case of read speech, such an average might be more informative because all people read the same text and, therefore, there is less variance associated with the content of what is being said. However, the performance gap is bridged after the application of the GA. This seems to suggest that the GA makes the CDMA effective at emphasizing depression-relevant information irrespective of the type of speech being used.

The main limitation of the work in this chapter is that the aggregation approach corresponds to the sum rule (Kittler et al., 1998). While being one of the methodologies most commonly applied for the combination of multiple classifiers, it is still possible to use more recent techniques designed, e.g., to combine multiple modalities. These methodologies include, e.g., joint and coordinated representations based on neural networks (see (Baltrusaitis et al., 2019) for an extensive survey). Furthermore, the GA can be difficult to extend, at least in its current version, to cases in which there are more than two data streams. However, despite these limitations, the CDMA has the major advantage that it can be applied to other problems in which there are two types of data that share a common property (depression or the lack of it in the case of this chapter) and that can be represented as sequences of feature vectors. Examples include, e.g., the audio and video streams of recordings to be categorized, multiple physiological signals collected for the same person, etc.

In view of a possible clinical application, it is important to observe that the performance of the system is comparable to those of *General Practitioners* (GP), i.e., doctors that are not specialized in psychiatry, but are typically the first line of intervention against the pathology. In fact, according to a meta-analysis of the literature (Mitchell et al., 2009), the accuracy of GPs ranges between 57.9% and 73.1%, while the average accuracy of the CDMA is 92.7%. On the other hand, the data used for the experiments were collected in real-world conditions (the room where patients meet their doctors), but still as a part of a controlled experiment. Furthermore, it is possible that in clinical practice, people refuse to be recorded or, especially when depression is severe, patients are not in the condition to perform the tasks at the core of the experiments (reading and answering questions). In this respect, the comparison above provides an indication, but it does not necessarily take into account all difficulties that might emerge in clinical practice.

From a technological point of view, the main direction for future work is to address the limitations mentioned earlier. In particular, the extension of the approach to more than two data streams will pave the way towards the inclusion of other modalities such as, e.g., the automatic transcription of spontaneous speech or, possibly, the videos of the experiment participants (this is not available in the data used for this thesis, but it is at disposition in other depression related corpora). Whenever the new streams of data provide information alternative to the others, there is a possibility to obtain further improvements, whether the problem is depression detection or any other type of classification.

Chapter 8

Conclusions and Future Work

8.1 Conclusions

In this thesis, we focus on speech-based depression detection. We propose a comprehensive methodology that consists of two main aspects, first, the introduction of effective markers, derived from both temporal and acoustic properties, which assist in the identification of depressed speakers; and second, the development and implementation of advanced deep learning models specifically designed for the detection of depression. Specifically, we hypothesize that measurable traces of depression can be found in speech, which can be discerned through speech properties and emphasized using attention mechanisms to differentiate between depressed and non-depressed speakers.

Firstly, we conducted a comprehensive survey of the currently available speech datasets for depression detection. Through this analysis, we identified several potential issues that may arise when using these corpora for depression detection:

- The distribution of depressed patients and controls participants might be unbalanced, or the overall number of participants may be insufficient for robust analysis;
- The dataset labels may be derived from questionnaires, which are inherently subjective assessments and can be prone to mislabeling;
- The speech data in the datasets may be limited in variety, with only one type of speech available.

To address these challenges and improve the quality of research in this field, we introduce a new, publicly available benchmark (Tao, Esposito, & Vinciarelli, 2023) in Chapter 3. Our proposed corpus offers several advantages that directly address the aforementioned issues:

- The distribution of depressed individuals and controls participants within the dataset is well-balanced, ensuring that both groups are adequately represented for analysis.

- This dataset is the largest publicly available speech corpus for depression detection with diagnosis by psychiatrists.
- To provide a more comprehensive understanding of depressive speech patterns, our corpus includes two different types of speech data, i.e., read speech and spontaneous speech.

The availability of this public corpus not only offers valuable resources for our investigation into depression detection in this thesis, but also presents numerous opportunities for other researchers to broaden their research interests in the area of automatic depression detection.

Secondly, from the perspective of cognitive neuroscience, we observe that depression changes the structure and function of the brain, and these alterations lead to changed speech processing mechanisms in the human brain, as well as in the muscles control responsible for speech production. This results in speech that differs from that of healthy individuals. This line of reasoning forms the foundation for our proposal of speech duration and silence as possible markers for speech-based depression detection (Tao et al., 2020), as discussed in Chapter 4. To rigorously evaluate the effectiveness of these proposed speech markers, we utilize techniques from computational linguistics (Schuller & Batliner, 2013) and social signal processing (Vinciarelli et al., 2009), and compare these markers to widely-used speech features, such as RMS energy, MFCC, and others. Finally, we assess their performance in both read and spontaneous speech. From a performance standpoint, we observe that incorporating speech duration and silence into our analysis leads to improved recognition accuracy. From a statistics standpoint, we find that these speech markers exhibit distinct differences between depressed and control groups, with speakers in the former group experiencing longer and more frequent periods of silence. This observation further supports our hypothesis, suggesting that individuals with depression require more time to process neural signals in order to produce speech.

In addition to the observed differences between depressed individuals and controls in the temporal aspect of speech, numerous acoustic features, such as MFCC and F0, also exhibit differences between the two groups. Given that speech is inherently a sequential signal, the relationships among acoustic features are expected to change over time. This observation motivates us to propose the use of correlation matrices of speech properties for identifying depressed speakers in Chapter 5. To rigorously evaluate the effectiveness of employing feature correlation matrices, we assess the performance across both read and spontaneous speech using various models. Experimental results demonstrate that the use of feature correlation matrices is more effective in detecting depressed speakers than the use of feature vectors. This finding remains robust across different models and both types of speech. Furthermore, our results suggest that feature correlation matrices introduce additional information, which exhibits greater variation in depressed individuals. Consequently, feature correlation matrices can be considered as potential markers for depression detection. These observations lend further support to our hypothesis, suggesting that depressed individuals experience changes in the control of muscles associated with speech production, ultimately leading to alterations in speech patterns.

The above sections mainly concentrate on the identification of speech markers in both temporal and acoustic dimensions for the purpose of depression detection. The findings indicate that these markers not only exhibit discernible differences between the speech of depressed and non-depressed individuals, but also serve to enhance the accuracy of detection. This observation supports our thesis statement that depression leaves measurable traces in both temporal and acoustic aspects of speech, which can subsequently be employed to differentiate between depressed and non-depressed speakers.

However, it is worth noting that there could be factors that impede the accurate modeling of depression-relevant information. For example, the inclusion of depression-irrelevant data, such as noise, could potentially result in erroneous outcomes. Therefore, this motivates us to use more sophisticated deep learning approaches for depression detection.

Regarding deep learning, we first analyzed the challenges of applying deep learning algorithms for speech-based depression detection. The major challenges come from weakly labeled speech signals and the equal treatment of all signals. To address the problem, we proposed a novel deep learning algorithm, namely MLA, for depression detection in Chapter 6. The core concept of our proposed MLA algorithm is to employ the attention mechanism, enabling the emphasis of depression-relevant information within short local frames. To thoroughly assess the effectiveness of the MLA, we evaluate its performance across both read and spontaneous speech. Our experimental results indicate that the MLA not only enhances recognition performance but also elevates efficiency and confidence in prediction. These findings are consistent across read speech and spontaneous speech. Remarkably, these improvements are consistent across both read and spontaneous speech, lending support to our hypothesis. This evidence suggests that the utilization of deep learning approaches, such as the MLA, allows for the highlighting of depression-relevant information, ultimately leading to more accurate and reliable depression detection using speech analysis.

Based on MLA, we further developed CDMA in Chapter 7 to facilitate the interaction between different types of speech for depression detection. The main limitation of MLA lies in its ability to only emphasize depression-relevant information within a single type of speech, thus restricting the effective combination of multiple speech types for depression detection. To tackle this issue, we introduced CDMA, which employs multiple attention mechanisms to emphasize depression-relevant information across various levels, fostering interaction between different types of speech for depression detection. Our experimental results suggest that the implementation of CDMA enables the exploration of emphasizing depression-relevant information in read and spontaneous speech, leading to an improvement in performance compared to the MLA approach. This observation supports our hypothesis that the traces of depression left in the speech include information specific to the type of speech and information common in speech.

In conclusion, the primary focus of Chapter 6 and Chapter 7 is the development of deep learning approaches for depression detection. The findings support our thesis statement that

leveraging attention mechanisms based on deep learning can lead to further advancements in performance for depression detection tasks.

8.2 Limitations

We introduce established corpus and approaches for automatic depression detection, but there are still some limitations.

The Androids Corpus has some limitations that should be considered. One of the main limitations is that it is not multimodal, but it is possible to transcribe it manually or automatically and still develop multimodal approaches based on language and paralanguage. In addition, the task order effect may be one of the limitations. Another limitation is that the dataset does not allow for an investigation of how depression progresses over time, as each participant is recorded only once. The data was collected in the clinical environment, and the effect of noise has not been considered. Although we employed psychiatrists to annotate the ground truth, the speech of participants may be affected by other uncontrollable individual differences, such as potential disorientation, exhaustion, etc. However, to the best of our knowledge, there are currently no publicly available corpora that address this problem. Despite these limitations, the dataset has the advantage of being recorded in real-world settings (hospitals) using standard laptop microphones and the distribution of gender is in line with the distribution in the real world. These advantages make it representative of the conditions in which depression detection is typically used.

One limitation of the findings related to speech markers is that they are exclusively based on Italian data, therefore, their acoustic and phonetic content might differ in other languages. The feature set employed in this thesis is standard, facilitating fair comparisons, however, there is a possibility of achieving improved results with a more refined set of features. For CDMA, there is a requirement to record two types of speech. Consequently, an interviewer is necessary for this process, but it does not necessarily need to be a human interviewer.

8.3 Future Work

In this section, we explore potential future directions for the field of automatic depression detection.

- **Applying attention mechanisms to silence:** In Chapter 4, we explored the benefits of utilizing silence as a marker for detecting depression. Our results indicated that individuals living with depression tend to manifest longer and more frequent silence. However, it should be noted that these variations in silence represent low-level descriptors, i.e., calculations based on entire recordings. Little is known about the silence changing over time.

In this study (Farruque et al., 2022), an innovative approach was proposed, which involves modeling the temporal aspects of silence within the textual content of social media as a method for detecting depression. This suggests that a similar technique could be utilized in speech, enabling the detection of depressed speakers through the modeling of temporal silence. This approach would facilitate an in-depth investigation into the temporal aspects of silence and how patterns of silence may shift in the speech of individuals with depression. Furthermore, the integration of attention mechanisms within this temporal silence model could underscore the significance of this "no activity" behavior. This, in turn, could provide a promising new direction in the field of depression detection, offering a novel way to highlight and analyze these periods of silence.

- **Combining feature vectors and feature correlation matrices for depression detection:** In Chapter 5, we explored the potential advantages of utilizing feature correlation matrices to detect depression speakers. Our findings indicate a greater degree of variation within the feature correlation matrices associated with those diagnosed with depression, suggesting that these matrices could serve as a valuable marker in depression detection. However, these correlation matrices are derived from feature vectors. From this perspective, it is possible to integrate feature vectors and feature correlation matrices for depression detection. For instance, such a relationship can be characterized by graph neural networks and used for recognition (K. Xu et al., 2018). Alternatively, a complementary approach would be to concatenate both the feature vectors and the feature correlation matrices, subsequently feeding them into deep networks for recognition. This integration would allow us to consider a wider spectrum of information, encompassing not only the speech features represented by feature vectors, but also the relationships between these features as encapsulated by the feature correlation matrices. This represents a promising direction for future research in the field of depression detection.

- **Interpretable deep learning models:** In Chapter 6 and Chapter 7, we have introduced leveraging attention mechanisms as a means to emphasize depression-relevant information. Our findings suggest that these approaches can significantly enhance the efficiency and performance of detecting depression. However, there is still a certain level of opacity in terms of the decision-making process of MLA and CDMA. This lack of transparency can be a potential barrier to fully understanding and utilizing the insights of these attention mechanisms, limiting the development of more efficient approaches for depression detection in the future. Therefore, one promising direction for future research could lie in the development of more interpretable deep learning models that can provide more clear explanations in the recognition. Another direction worth exploring is the potent visualization of attention mechanisms. Drawing inspiration from the field of computer vision, attention visualization offers an explicit depiction of the focus areas within images or videos (Ge

et al., 2023). This facilitates a direct understanding of how the model assesses and selects certain areas over others. Therefore, applying such visualization to speech analysis could significantly enhance our understanding of the decision-making process of the model, e.g., a direct explanation of which words or features in the speech are important for depression detection.

- **Recognizing depression with multimodal data:** In this thesis, our primary emphasis lies on the utilization of speech as a means to detect depression. Our results demonstrate that the use of speech is able to detect depressed speakers. However, the approaches explored in this thesis are not based on multimodal analysis. Recently, there has been a burgeoning interest in the application of multimodal strategies for depression detection (Aloshban et al., 2020, 2021). In light of this, a promising direction for future research involves the transcription of speech into textual data. This is particularly pertinent for spontaneous speech, which may provide more information on the mental state of speakers (Kiss & Vicsi, 2017a). As a result, this will not only augment the corpus at our disposal but also empower researchers to incorporate a linguistic perspective when formulating multimodal approaches for depression detection.

- **Ethical Considerations:** As the prevalence of automatic depression detection increases, it is crucial to consider the ethical implications of such technology. Currently, privacy issues have been raised by researchers as potential challenges in the field of automatic depression detection, e.g., whether depressed speakers can be identified, etc (Bowie-DaBreo et al., 2022). One of the future directions is to pay attention to privacy protection and avoid the potential misuse of the technology.

8.4 Concluding Remarks

This thesis has tackled the demanding task of implementing automatic depression detection based on speech. Specifically, it has made contributions to the field by integrating potential speech markers and deep learning methodologies with attention mechanisms. We have shown that speech duration, silence and feature correlation matrices can effectively enhance the performance in recognition. The incorporation of attention mechanisms within deep learning approaches has also proven beneficial in improving both detection performance and efficiency. However, there are still some interesting topics and challenges in this research field. We have highlighted the potential future directions for addressing these challenges in Section 8.3. We firmly believe that the continued application of speech-based and deep learning methodologies will be instrumental in the development of automatic depression detection.

References

Abdullaev, Y., Kennedy, B. L., & Tasman, A. (2002). Changes in neural circuitry of language before and after treatment of major depression. *Human Brain Mapping*, *17*(3), 156–167.

Abhang, P. A., Gawali, B. W., & Mehrotra, S. C. (2016). Technical aspects of brain rhythms and speech parameters. *Introduction to EEG-and Speech-based Emotion Recognition*, 51–79.

Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Umar, A. M., Linus, O. U., ... Kiru, M. U. (2019). Comprehensive review of artificial neural network applications to pattern recognition. *IEEE Access*, *7*, 158820–158846.

Albuquerque, L., Valente, A. R. S., Teixeira, A., Figueiredo, D., Sa-Couto, P., & Oliveira, C. (2021). Association between acoustic speech features and non-severe levels of anxiety and depression symptoms across lifespan. *PloS One*, *16*(4), e0248842.

Alexander, M. P., Benson, D. F., & Stuss, D. T. (1989). Frontal lobes and language. *Brain and Language*, *37*(4), 656–691.

Alghowinem, S., Goecke, R., Epps, J., Wagner, M., & Cohn, J. F. (2016). Cross-cultural depression recognition from vocal biomarkers. *Proceedings of INTERSPEECH*, 1943–1947.

Alghowinem, S., Goecke, R., Wagner, M., Epps, J., Breakspear, M., & Parker, G. (2012). From joyous to clinically depressed: Mood detection using spontaneous speech. *Proceedings of FLAIRS Conference*, *19*.

Alghowinem, S., Goecke, R., Wagner, M., Epps, J., Breakspear, M., & Parker, G. (2013). Detecting depression: a comparison between spontaneous and read speech. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 7547–7551.

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

Alghowinem, S., Goecke, R., Wagner, M., Parkerx, G., & Breakspear, M. (2013). Head pose and movement analysis as an indicator of depression. *Proceedings of the IEEE International Conference on Affective Computing and Intelligent Interaction*, 283–288.

Al Hanai, T., Ghassemi, M. M., & Glass, J. R. (2018). Detecting depression with audio/text sequence modeling of interviews. *Proceedings of INTERSPEECH*, 1716–1720.

Almaghrabi, S. A., Clark, S. R., & Baumert, M. (2023). Bio-acoustic features of depression: A review. *Biomedical Signal Processing and Control*, *85*, 105020.

Aloshban, N., Esposito, A., & Vinciarelli, A. (2020). Detecting depression in less than 10 seconds: Impact of speaking time on depression detection sensitivity. *Proceedings of the International Conference on Multimodal Interaction*, 79–87.

Aloshban, N., Esposito, A., & Vinciarelli, A. (2021). Language or paralanguage, this is the problem: Comparing depressed and non-depressed speakers through the analysis of gated multimodal units. *Proceedings of INTERSPEECH*, 2496–2500.

Aloshban, N., Esposito, A., & Vinciarelli, A. (2022). What you say or how you say it? depression detection through joint modeling of linguistic and acoustic aspects of speech. *Cognitive Computation*, *14*(5), 1585–1598.

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

Alsarrani, R., Esposito, A., & Vinciarelli, A. (2022). Thin slices of depression: Improving depression detection performance through data segmentation. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 6257–6261.

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, *111*(2), 256-274.

Andrade, L., Caraveo-Anduaga, J., Berglund, P., Bijl, R., De Graaf, R., Vollebergh, W., ... Wittchen, H.-U. (2003). The epidemiology of major depressive episodes: Results from the International Consortium of Psychiatric Epidemiology (ICPE) surveys. *International Journal of Methods in Psychiatric Research*, *12*(1), 3–21.

APA, A. P. A. (2013). Diagnostic and statistical manual of mental disorders. *The American Psychiatric Association*.

Backes, H., Dietsche, B., Nagels, A., Stratmann, M., Konrad, C., Kircher, T., & Krug, A. (2014). Increased neural activity during overt and continuous semantic verbal fluency in major depression: mainly a failure to deactivate. *European Archives of Psychiatry and Clinical Neuroscience*, *264*, 631–645.

Baltrusaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*(2), 423–443.

Basu, S., Chakraborty, J., Bag, A., & Aftabuddin, M. (2017). A review on emotion recognition using speech. *Proceedings of the International Conference on Inventive Communication and Computational Technologies*, 109–114.

Beck, A. T. (2008). The evolution of the cognitive model of depression and its neurobiological correlates. *American Journal of Psychiatry*, *165*(8), 969–977.

Beck, A. T., Steer, R. A., & Brown, G. (1996). Beck depression inventory–ii. *Psychological Assessment*.

Bedny, M., Hulbert, J. C., & Thompson-Schill, S. L. (2007). Understanding words in context: the role of broca's area in word comprehension. *Brain Research*, *1146*, 101–114.

Beevers, C. G., Clasen, P., Stice, E., & Schnyer, D. (2010). Depression symptoms and cognitive control of emotion cues: a functional magnetic resonance imaging study. *Neuroscience*, *167*(1), 97–103.

Benesty, J., Sondhi, M. M., Huang, Y., et al. (2008). *Springer handbook of speech processing* (Vol. 1). Springer.

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, *5*(2), 157–166.

Bernard, J. D., Baddeley, J. L., Rodriguez, B. F., & Burke, P. A. (2016). Depression, language, and affect: an examination of the influence of baseline depression and affect induction on language. *Journal of Language and Social Psychology*, *35*(3), 317–326.

Binder, J. R., McKiernan, K. A., Parsons, M. E., Westbury, C. F., Possing, E. T., Kaufman, J. N., & Buchanan, L. (2003). Neural correlates of lexical access during visual word recognition. *Journal of Cognitive Neuroscience*, *15*(3), 372–393.

Boersma, P., et al. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proceedings of the Institute of Phonetic Sciences*, *17*(1193), 97–110.

Bowie-DaBreo, D., Sas, C., Iles-Smith, H., & Sünram-Lea, S. (2022). User perspectives and ethical experiences of apps for depression: A qualitative analysis of user reviews. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–24.

Brewer, R., Pierce, C., Upadhyay, P., & Park, L. (2022). An empirical study of older adult's voice assistant use for health information seeking. *ACM Transactions on Interactive Intelligent Systems*, *12*(2), 1–32.

Brigham, E. O. (1988). *The fast fourier transform and its applications*. Prentice-Hall, Inc.

Bueno-Notivol, J., Gracia-García, P., Olaya, B., Lasheras, I., López-Antón, R., & Santabárbara, J. (2021). Prevalence of depression during the covid-19 outbreak: A meta-analysis of community-based studies. *International Journal of Clinical and Health Psychology*, *21*(1), 100196.

Burton, C., McKinstry, B., Tătar, A. S., Serrano-Blanco, A., Pagliari, C., & Wolters, M. (2013). Activity monitoring in patients with depression: a systematic review. *Journal of Affective Disorders*, *145*(1), 21–28.

Bush, G., Luu, P., & Posner, M. I. (2000). Cognitive and emotional influences in anterior cingulate cortex. *Trends in Cognitive Sciences*, *4*(6), 215–222.

Cabeza, R., & Nyberg, L. (2000). Imaging cognition ii: An empirical review of 275 pet and fmri studies. *Journal of Cognitive Neuroscience*, *12*(1), 1–47.

Cai, C., Niu, M., Liu, B., Tao, J., & Liu, X. (2021). Tdca-net: Time-domain channel attention network for depression detection. *Proceedings of INTERSPEECH*, 2511–2515.

Cai, H., Gao, Y., Sun, S., Li, N., Tian, F., Xiao, H., ... others (2020). Modma dataset: a multi-modal open dataset for mental-disorder analysis. *arXiv preprint arXiv:2002.09283*.

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

Carding, P. N., Wilson, J. A., MacKenzie, K., & Deary, I. J. (2009). Measuring voice outcomes: state of the science review. *The Journal of Laryngology & Otology*, *123*(8), 823–829.

Chai, T., & 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.

Cheng, W., Rolls, E. T., Qiu, J., Liu, W., Tang, Y., Huang, C.-C., ... others (2016). Medial reward and lateral non-reward orbitofrontal cortex circuits change in opposite directions in depression. *Brain*, *139*(12), 3296–3309.

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

Corbetta, M., Akbudak, E., Conturo, T. E., Snyder, A. Z., Ollinger, J. M., Drury, H. A., ... others (1998). A common network of functional areas for attention and eye movements. *Neuron*, *21*(4), 761–773.

Costafreda, S. G., Brammer, M. J., David, A. S., & Fu, C. H. (2008). Predictors of amygdala activation during the processing of emotional stimuli: a meta-analysis of 385 pet and fmri studies. *Brain Research Reviews*, *58*(1), 57–70.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press.

Cummins, N., Epps, J., & Ambikairajah, E. (2013). Spectro-temporal analysis of speech affected by depression and psychomotor retardation. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 7542–7546.

Cummins, N., Epps, J., Breakspear, M., & Goecke, R. (2011). An investigation of depressed speech detection: Features and normalization. *Proceedings of INTERSPEECH*.

Cummins, N., Epps, J., Sethu, V., Breakspear, M., & Goecke, R. (2013). Modeling spectral variability for the classification of depressed speech. *Proceedings of INTERSPEECH*, 857–861.

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

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

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

Cummins, N., Sethu, V., Epps, J., Williamson, J., Quatieri, T., & Krajewski, J. (2020). Generalized two-stage rank regression framework for depression score prediction from speech. *IEEE Transactions on Affective Computing*, *11*(2), 272-283.

D'Arcy, R. C., Connolly, J. F., Service, E., Hawco, C. S., & Houlihan, M. E. (2004). Separating phonological and semantic processing in auditory sentence processing: A high-resolution event-related brain potential study. *Human Brain Mapping*, *22*(1), 40–51.

Davidson, R. J., Pizzagalli, D., Nitschke, J. B., & Putnam, K. (2002). Depression: perspectives from affective neuroscience. *Annual Review of Psychology*, *53*(1), 545–574.

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2010). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(4), 788–798.

Del Arco, A., & Mora, F. (2008). Prefrontal cortex–nucleus accumbens interaction: in vivo modulation by dopamine and glutamate in the prefrontal cortex. *Pharmacology Biochemistry and Behavior*, *90*(2), 226–235.

Demitrack, M. A., Faries, D., Herrera, J. M., DeBrota, D. J., & Potter, W. Z. (1998). The problem of measurement error in multisite clinical trials. *Psychopharmacology Bulletin*, *34*(1), 19.

DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., … others (2014). Simsensei kiosk: A virtual human interviewer for healthcare decision support. *Proceedings of the International Conference on Autonomous Agents and Multi-agent Systems*, 1061–1068.

Disner, S. G., Beevers, C. G., Haigh, E. A., & Beck, A. T. (2011). Neural mechanisms of the cognitive model of depression. *Nature Reviews Neuroscience*, *12*(8), 467–477.

Dobson, K. S. (1989). A meta-analysis of the efficacy of cognitive therapy for depression. *Journal of Consulting and Clinical Psychology*, *57*(3), 414.

Dong, Y., & Yang, X. (2021). A hierarchical depression detection model based on vocal and emotional cues. *Neurocomputing*, *441*, 279–290.

Drevets, W. C. (2001). Neuroimaging and neuropathological studies of depression: implications for the cognitive-emotional features of mood disorders. *Current Opinion in Neurobiology*, *11*(2), 240–249.

Dronkers, N. F., Plaisant, O., Iba-Zizen, M. T., & Cabanis, E. A. (2007). Paul broca's historic cases: high resolution mr imaging of the brains of leborgne and lelong. *Brain*, *130*(5), 1432–1441.

Du, M., Liu, S., Wang, T., Zhang, W., Ke, Y., Chen, L., & Ming, D. (2023). Depression recognition using a proposed speech chain model fusing speech production and perception features. *Journal of Affective Disorders*, *323*, 299–308.

Du, M., Zhang, W., Wang, T., Liu, S., & Ming, D. (2022). An automatic depression recognition method from spontaneous pronunciation using machine learning. *Proceedings of the 9th International Conference on Biomedical and Bioinformatics Engineering*, 133–139.

Dumpala, S. H., Rempel, S., Dikaios, K., Sajjadian, M., Uher, R., & Oore, S. (2021). Estimating severity of depression from acoustic features and embeddings of natural speech. *Proceedings*

of the IEEE International Conference on Acoustics, Speech and Signal Processing, 7278–7282.

Dumpala, S. H., Rodriguez, S., Rempel, S., Sajjadian, M., Uher, R., & Oore, S. (2022). Detecting depression with a temporal context of speaker embeddings. *Proceedings of the AAAI SAS*.

Egas-López, J. V., Kiss, G., Sztahó, D., & Gosztolya, G. (2022). Automatic assessment of the degree of clinical depression from speech using x-vectors. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 8502–8506.

Elfgren, C. I., & Risberg, J. (1998). Lateralized frontal blood flow increases during fluency tasks: influence of cognitive strategy. *Neuropsychologia*, *36*(6), 505–512.

Espinola, C. W., Gomes, J. C., Pereira, J. M. S., & dos Santos, W. P. (2021). Detection of major depressive disorder using vocal acoustic analysis and machine learning—an exploratory study. *Research on Biomedical Engineering*, *37*, 53–64.

Esposito, A., Esposito, A. M., Likforman-Sulem, L., Maldonato, M. N., & Vinciarelli, A. (2016). On the significance of speech pauses in depressive disorders: results on read and spontaneous narratives. *Recent Advances in Nonlinear Speech Processing*, 73–82.

Eugène, F., Joormann, J., Cooney, R. E., Atlas, L. Y., & Gotlib, I. H. (2010). Neural correlates of inhibitory deficits in depression. *Psychiatry Research: Neuroimaging*, *181*(1), 30–35.

Evans-Lacko, S., Aguilar-Gaxiola, S., Al-Hamzawi, A., Alonso, J., Benjet, C., Bruffaerts, R., ... others (2018). Socio-economic variations in the mental health treatment gap for people with anxiety, mood, and substance use disorders: results from the who world mental health (wmh) surveys. *Psychological Medicine*, *48*(9), 1560–1571.

Eyben, F., Weninger, F., Gross, F., & Schuller, B. (2013). Recent developments in Opensmile, the Munich open-source multimedia feature extractor. *Proceedings of the ACM International Conference on Multimedia*, 835–838.

Faisal-Cury, A., Ziebold, C., de Oliveira Rodrigues, D. M., & Matijasevich, A. (2022). Depression underdiagnosis: Prevalence and associated factors. a population-based study. *Journal of Psychiatric Research*, *151*, 157–165.

Fales, C. L., Barch, D. M., Rundle, M. M., Mintun, M. A., Snyder, A. Z., Cohen, J. D., ... Sheline, Y. I. (2008). Altered emotional interference processing in affective and cognitive-control brain circuitry in major depression. *Biological Psychiatry*, *63*(4), 377–384.

Farruque, N., Goebel, R., Sivapalan, S., & Zaïane, O. R. (2022). Deep temporal modelling of clinical depression through social media text. *arXiv preprint arXiv:2211.07717*.

Fast, L. A., & Funder, D. C. (2010). Gender differences in the correlates of self-referent word use: Authority, entitlement, and depressive symptoms. *Journal of Personality*, *78*(1), 313–338.

Fava, M., Evins, A. E., Dorer, D. J., & Schoenfeld, D. A. (2003). The problem of the placebo response in clinical trials for psychiatric disorders: culprits, possible remedies, and a novel study design approach. *Psychotherapy and Psychosomatics*, *72*(3), 115–127.

Fay, D., & Cutler, A. (1977). Malapropisms and the structure of the mental lexicon. *Linguistic Inquiry*, *8*(3), 505–520.

Flint, A. J., Black, S. E., Campbell-Taylor, I., Gailey, G. F., & Levinton, C. (1993). Abnormal speech articulation, psychomotor retardation, and subcortical dysfunction in major depression. *Journal of Psychiatric Research*, *27*(3), 309–319.

France, D. J., Shiavi, R. G., Silverman, S., Silverman, M., & Wilkes, M. (2000). Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE Transactions on Biomedical Engineering*, *47*(7), 829–837.

Friston, K. J. (2011). Functional and effective connectivity: a review. *Brain Connectivity*, *1*(1), 13–36.

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-P., Frith, C. D., & Frackowiak, R. S. (1994). Statistical parametric maps in functional imaging: a general linear approach. *Human Brain Mapping*, *2*(4), 189–210.

Gaillard, W. D., Hertz-Pannier, L., Mott, S. H., Barnett, A. S., LeBihan, D., & Theodore, W. H. (2000). Functional anatomy of cognitive development: fmri of verbal fluency in children and adults. *Neurology*, *54*(1), 180–180.

Ge, X., Chen, F., Xu, S., Tao, F., & Jose, J. M. (2023). Cross-modal semantic enhanced interaction for image-sentence retrieval. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1022–1031.

Ge, X., Wang, P., Han, H., Jose, J. M., Ji, Z., Wu, Z., & Liu, X. (2021). Local global relational network for facial action units recognition. *Proceedings of the 16th IEEE International Conference on Automatic Face and Gesture Recognition*, 01–08.

Gers, F. A., Schmidhuber, J., & Cummins, F. (2000). Learning to forget: Continual prediction with lstm. *Neural Computation*, *12*(10), 2451–2471.

Gobl, C., & Chasaide, A. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech Communication*, *40*(1-2), 189–212.

Gotlib, I. H., & Hamilton, J. P. (2008). Neuroimaging and depression: Current status and unresolved issues. *Current Directions in Psychological Science*, *17*(2), 159–163.

Gotlib, I. H., & Hammen, C. L. (2008). *Handbook of depression*. Guilford Press.

Gotlib, I. H., Krasnoperova, E., Yue, D. N., & Joormann, J. (2004). Attentional biases for negative interpersonal stimuli in clinical depression. *Journal of Abnormal Psychology*, *113*(1), 127.

Gratch, J., Artstein, R., Lucas, G., Stratou, G., Scherer, S., Nazarian, A., ... others (2014). *The distress analysis interview corpus of human and computer interviews* (Tech. Rep.). University of Southern California Los Angeles.

Greenberg, P. E., Fournier, A.-A., Sisitsky, T., Pike, C. T., & Kessler, R. C. (2015). The economic burden of adults with major depressive disorder in the united states (2005 and 2010). *Journal of Clinical Psychiatry*, *76*(2), 5356.

Greenberg, P. E., Fournier, A.-A., Sisitsky, T., Simes, M., Berman, R., Koenigsberg, S. H., & Kessler, R. C. (2021). The economic burden of adults with major depressive disorder in the united states (2010 and 2018). *Pharmacoeconomics*, *39*(6), 653–665.

Greenberg, P. E., Kessler, R. C., Birnbaum, H. G., Leong, S. A., Lowe, S. W., Berglund, P. A., & Corey-Lisle, P. K. (2003). The economic burden of depression in the united states: how did it change between 1990 and 2000? *Journal of Clinical Psychiatry*, *64*(12), 1465–1475.

Greenberg, P. E., Stiglin, L. E., Finkelstein, S. N., & Berndt, E. R. (1993). The economic burden of depression in 1990. *Journal of Clinical Psychiatry*.

Greenberg, R. P., Bornstein, R. F., Greenberg, M. D., & Fisher, S. (1992). A meta-analysis of antidepressant outcome under" blinder" conditions. *Journal of Consulting and Clinical Psychology*, *60*(5), 664.

Guidi, A., Vanello, N., Bertschy, G., Gentili, C., Landini, L., & Scilingo, E. P. (2015). Automatic analysis of speech f0 contour for the characterization of mood changes in bipolar patients. *Biomedical Signal Processing and Control*, *17*, 29–37.

Haji, T., Horiguchi, S., Baer, T., & Gould, W. J. (1986). Frequency and amplitude perturbation analysis of electroglottograph during sustained phonation. *The Journal of the Acoustical Society of America*, *80*(1), 58–62.

Hamilton, M. (1986). The hamilton rating scale for depression. *Assessment of Depression*, 143–152.

Harati, A., Shriberg, E., Rutowski, T., Chlebek, P., Lu, Y., & Oliveira, R. (2021). Speech-based depression prediction using encoder-weight-only transfer learning and a large corpus. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 7273–7277.

Hasin, D. S., Sarvet, A. L., Meyers, J. L., Saha, T. D., Ruan, W. J., Stohl, M., & Grant, B. F. (2018). Epidemiology of adult dsm-5 major depressive disorder and its specifiers in the united states. *JAMA Psychiatry*, *75*(4), 336–346.

He, L., & Cao, C. (2018). Automated depression analysis using convolutional neural networks from speech. *Journal of Biomedical Informatics*, *83*, 103–111.

He, L., Niu, M., Tiwari, P., Marttinen, P., Su, R., Jiang, J., ... others (2022). Deep learning for depression recognition with audiovisual cues: A review. *Information Fusion*, *80*, 56–86.

Helfer, B. S., Quatieri, T. F., Williamson, J. R., Mehta, D. D., Horwitz, R., & Yu, B. (2013). Classification of depression state based on articulatory precision. *Proceedings of INTERSPEECH*, 2172–2176.

Heller, A. S., Johnstone, T., Shackman, A. J., Light, S. N., Peterson, M. J., Kolden, G. G., ... Davidson, R. J. (2009). Reduced capacity to sustain positive emotion in major depression reflects diminished maintenance of fronto-striatal brain activation. *Proceedings of the National Academy of Sciences*, *106*(52), 22445–22450.

Heller, W., & Nitscke, J. B. (1997). Regional brain activity in emotion: A framework for understanding cognition in depresion. *Cognition & Emotion*, *11*(5-6), 637–661.

Highland, D., & Zhou, G. (2022). A review of detection techniques for depression and bipolar disorder. *Smart Health*, 100282.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Holahan, C. J., & Moos, R. H. (1987). Personal and contextual determinants of coping strategies. *Journal of Personality and Social Psychology*, *52*(5), 946.

Holahan, C. J., Moos, R. H., Holahan, C. K., Brennan, P. L., & Schutte, K. K. (2005). Stress generation, avoidance coping, and depressive symptoms: a 10-year model. *Journal of Consulting and Clinical Psychology*, *73*(4), 658.

Holtzman, N. S., et al. (2017). A meta-analysis of correlations between depression and first person singular pronoun use. *Journal of Research in Personality*, *68*, 63–68.

Hönig, F., Batliner, A., Nöth, E., Schnieder, S., & Krajewski, J. (2014). Automatic modelling of depressed speech: relevant features and relevance of gender.

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

Howell, D. (2009). *Statistical methods for psychology*. Cengage Learning.

Huang, Z., Epps, J., & Joachim, D. (2019a). Investigation of speech landmark patterns for depression detection. *IEEE Transactions on Affective Computing*.

Huang, Z., Epps, J., & Joachim, D. (2019b). Speech landmark bigrams for depression detection from naturalistic smartphone speech. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 5856–5860.

Huang, Z., Epps, J., & Joachim, D. (2020). Exploiting vocal tract coordination using dilated cnns for depression detection in naturalistic environments. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 6549–6553.

Huang, Z., Epps, J., Joachim, D., & Chen, M. (2018). Depression detection from short utterances via diverse smartphones in natural environmental conditions. *Proceedings of INTERSPEECH*, 3393–3397.

Huang, Z., Epps, J., Joachim, D., Stasak, B., Williamson, J. R., & Quatieri, T. F. (2020). Domain adaptation for enhancing speech-based depression detection in natural environmental conditions using dilated cnns. *Proceedings of INTERSPEECH*, 4561–4565.

Hunt, M., Auriemma, J., & Cashaw, A. C. (2003). Self-report bias and underreporting of depression on the bdi-ii. *Journal of personality assessment*, *80*(1), 26–30.

Hussenbocus, A. Y., Lech, M., & Allen, N. B. (2015). Statistical differences in speech acoustics of major depressed and non-depressed adolescents. *Proceedings of the 9th International Conference on Signal Processing and Communication Systems*, 1–7.

Ising, M., Horstmann, S., Kloiber, S., Lucae, S., Binder, E. B., Kern, N., ... Holsboer, F. (2007). Combined dexamethasone/corticotropin releasing hormone test predicts treatment response in major depression–a potential biomarker? *Biological Psychiatry*, *62*(1), 47–54.

Jarrold, W., Javitz, H. S., Krasnow, R., Peintner, B., Yeh, E., Swan, G. E., & Mehl, M. (2011). Depression and self-focused language in structured interviews with older men. *Psychological Reports*, *109*(2), 686–700.

Jia, Y., Liang, Y., & Zhu, T. (2019). An analysis of voice quality of chinese patients with depression. *Proceedings of the 22nd Conference of the Oriental COCOSDA International Committee for the Coordination and Standardisation of Speech Databases and Assessment Techniques*, 1–6.

Jia, Y., Liang, Y., & Zhu, T. (2020). An analysis of acoustic features in reading speech from chinese patients with depression. *Proceedings of the 23rd Conference of the Oriental COCOSDA International Committee for the Coordination and Standardisation of Speech Databases and Assessment Techniques*, 128–133.

Jiang, H., Hu, B., Liu, Z., Wang, G., Zhang, L., Li, X., & Kang, H. (2018). Detecting depression using an ensemble logistic regression model based on multiple speech features. *Computational and Mathematical Methods in Medicine*, *2018*.

Jiang, H., Hu, B., Liu, Z., Yan, L., Wang, T., Liu, F., ... Li, X. (2017). Investigation of different speech types and emotions for detecting depression using different classifiers. *Speech Communication*, *90*, 39–46.

Johnson, K. (2004). Massive reduction in conversational american english. *Spontaneous speech: Data and analysis. Proceedings of the 1st Session of the 10th International Symposium*, 29–54.

Kann, L., McManus, T., Harris, W. A., Shanklin, S. L., Flint, K. H., Queen, B., ... others (2018). Youth risk behavior surveillance—united states, 2017. *MMWR Surveillance Summaries*, *67*(8), 1.

Karam, Z. N., Provost, E. M., Singh, S., Montgomery, J., Archer, C., Harrington, G., & Mcinnis, M. G. (2014). Ecologically valid long-term mood monitoring of individuals with bipolar disorder using speech. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 4858–4862.

Kellough, J. L., Beevers, C. G., Ellis, A. J., & Wells, T. T. (2008). Time course of selective attention in clinically depressed young adults: An eye tracking study. *Behaviour Research and Therapy*, *46*(11), 1238–1243.

Kennison, S. M. (2013). *Introduction to language development*. Sage Publications.

Kenny, P., Boulianne, G., & Dumouchel, P. (2005). Eigenvoice modeling with sparse training data. *IEEE Transactions on Speech and Audio Processing*, *13*(3), 345–354.

Kent, R. D. (2000). Research on speech motor control and its disorders: A review and prospective. *Journal of Communication Disorders*, *33*(5), 391–428.

Kessler, R., Berglund, P., Demler, O., Jin, R., Koretz, D., Merikangas, K., ... Wang, P. (2003). The epidemiology of major depressive disorder: Results from the National Comorbidity Survey Replication (NCS-R). *Journal of the American Medical Association*, *289*(23), 3095–3105.

Kessler, R. C. (2003). Epidemiology of women and depression. *Journal of Affective Disorders*, *74*(1), 5–13.

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Koretz, D., Merikangas, K. R., ... Wang, P. S. (2003). The epidemiology of major depressive disorder: results from the national comorbidity survey replication (ncs-r). *JAMA*, *289*(23), 3095–3105.

Kessler, R. C., McGonagle, K. A., Zhao, S., Nelson, C. B., Hughes, M., Eshleman, S., ... Kendler, K. S. (1994). Lifetime and 12-month prevalence of dsm-iii-r psychiatric disorders in the united states: results from the national comorbidity survey. *Archives of General Psychiatry*, *51*(1), 8–19.

Khorram, S., Gideon, J., McInnis, M. G., & Provost, E. M. (2016). Recognition of depression in bipolar disorder: Leveraging cohort and person-specific knowledge. *Proceedings of INTERSPEECH*, 1215–1219.

Kießling, A. (1997). *Extraktion und klassifikation prosodischer merkmale in der automatischen sprachverarbeitung*. Shaker.

Kim, A. Y., Jang, E. H., Lee, S.-H., Choi, K.-Y., Park, J. G., & Shin, H.-C. (2023). Automatic depression detection using smartphone-based text-dependent speech signals: Deep convolutional neural network approach. *Journal of Medical Internet Research*, *25*, e34474.

Kim, S. H., & Hamann, S. (2007). Neural correlates of positive and negative emotion regulation. *Journal of Cognitive Neuroscience*, *19*(5), 776–798.

Kinnunen, T., & Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication*, *52*(1), 12–40.

Kiss, G., Tulics, M. G., Sztahó, D., Esposito, A., & Vicsi, K. (2016). Language independent detection possibilities of depression by speech. *Recent Advances in Nonlinear Speech Processing*, 103–114.

Kiss, G., & Vicsi, K. (2017a). Comparison of read and spontaneous speech in case of automatic detection of depression. *Proceedings of the 8th IEEE International Conference on Cognitive Infocommunications*, 000213–000218.

Kiss, G., & Vicsi, K. (2017b). Mono-and multi-lingual depression prediction based on speech processing. *International Journal of Speech Technology*, *20*, 919–935.

Kittler, J., Hatef, M., Duin, R. P., & Matas, J. (1998). On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(3), 226–239.

Klumpp, H., & Deldin, P. (2010). Review of brain functioning in depression for semantic processing and verbal fluency. *International Journal of Psychophysiology*, *75*(2), 77–85.

Koops, S., Brederoo, S. G., de Boer, J. N., Nadema, F. G., Voppel, A. E., & Sommer, I. E. (2023). Speech as a biomarker for depression. *CNS & Neurological Disorders-Drug Targets (Formerly Current Drug Targets-CNS & Neurological Disorders)*, *22*(2), 152–160.

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The phq-9: validity of a brief depression severity measure. *Journal of General Internal Medicine*, *16*(9), 606–613.

Larsen, M. E., Cummins, N., Boonstra, T. W., O'Dea, B., Tighe, J., Nicholas, J., ... Christensen, H. (2015). The use of technology in suicide prevention. *Proceedings of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 7316–7319.

LeDoux, J. (2007). The amygdala. *Current Biology*, *17*(20), R868–R874.

LeDoux, J. E. (2000). Emotion circuits in the brain. *Annual Review of Neuroscience*, *23*(1), 155–184.

Lee, B.-H., Kim, H., Park, S.-H., & Kim, Y.-K. (2007). Decreased plasma bdnf level in depressive patients. *Journal of Affective Disorders*, *101*(1-3), 239–244.

Li, C.-T., Lin, C.-P., Chou, K.-H., Chen, I.-Y., Hsieh, J.-C., Wu, C.-L., ... Su, T.-P. (2010). Structural and cognitive deficits in remitting and non-remitting recurrent depression: a voxel-based morphometric study. *Neuroimage*, *50*(1), 347–356.

Linsley, D., Shiebler, D., Eberhardt, S., & Serre, T. (2019). Learning what and where to attend. *Proceedings of the International Conference on Learning Representations*.

Liu, Z., Kang, H., Feng, L., & Zhang, L. (2017). Speech pause time: A potential biomarker for depression detection. *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine*, 2020–2025.

Liu, Z., Li, C., Gao, X., Wang, G., & Yang, J. (2017). Ensemble-based depression detection in speech. *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine*, 975–980.

Long, H., Guo, Z., Wu, X., Hu, B., Liu, Z., & Cai, H. (2017). Detecting depression in speech: Comparison and combination between different speech types. *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine*, 1052–1058.

Low, D. M., Bentley, K. H., & Ghosh, S. S. (2020). Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investigative Otolaryngology*, *5*(1), 96–116.

Low, L.-S. A., Maddage, N. C., Lech, M., Sheeber, L. B., & Allen, N. B. (2010). Detection of clinical depression in adolescents' speech during family interactions. *IEEE Transactions on Biomedical Engineering*, *58*(3), 574–586.

Lu, X., Shi, D., Liu, Y., & Yuan, J. (2021). Speech depression recognition based on attentional residual network. *Frontiers in Bioscience-Landmark*, *26*(12), 1746–1759.

Mallol-Ragolta, A., Zhao, Z., Stappen, L., Cummins, N., & Schuller, B. (2019). A hierarchical attention network-based approach for depression detection from transcribed clinical interviews. *Proceedings of INTERSPEECH*.

Mathers, C. D., & Loncar, D. (2006). Projections of global mortality and burden of disease from 2002 to 2030. *PLoS Medicine*, *3*(11), e442.

Mathews, A., & MacLeod, C. (2005). Cognitive vulnerability to emotional disorders. *Annual Review of Clinical Psychology*, *1*, 167–195.

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes*, *27*(7-8), 953–978.

McGinnis, E. W., Anderau, S. P., Hruschak, J., Gurchiek, R. D., Lopez-Duran, N. L., Fitzgerald, K., ... McGinnis, R. S. (2019). Giving voice to vulnerable children: machine learning analysis of speech detects anxiety and depression in early childhood. *IEEE Journal of Biomedical and Health Informatics*, *23*(6), 2294–2301.

McLaughlin, K. A. (2011). The public health impact of major depression: a call for interdisciplinary prevention efforts. *Prevention Science*, *12*, 361–371.

Meng, X.-L., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin*, *111*(1), 172.

Mitchell, A. J., Vaze, A., & Rao, S. (2009). Clinical diagnosis of depression in primary care: a meta-analysis. *The Lancet*, *374*(9690), 609–619.

Mitra, V., & Shriberg, E. (2015). Effects of feature type, learning algorithm and speaking style for depression detection from speech. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 4774–4778.

Mitra, V., Shriberg, E., Vergyri, D., Knoth, B., & Salomon, R. M. (2015). Cross-corpus depression prediction from speech. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 4769–4773.

Mitra, V., Tsiartas, A., & Shriberg, E. (2016). Noise and reverberation effects on depression detection from speech. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 5795–5799.

Mitterschiffthaler, M., Williams, S., Walsh, N., Cleare, A., Donaldson, C., Scott, J., & Fu, C. (2008). Neural basis of the emotional stroop interference effect in major depression. *Psychological Medicine*, *38*(2), 247–256.

Mojtabai, R., Olfson, M., & Han, B. (2016). National trends in the prevalence and treatment of depression in adolescents and young adults. *Pediatrics*, *138*(6).

Molendijk, M. L., Bamelis, L., van Emmerik, A. A., Arntz, A., Haringsma, R., & Spinhoven, P. (2010). Word use of outpatients with a personality disorder and concurrent or previous major depressive disorder. *Behaviour Research and Therapy*, *48*(1), 44–51.

Montgomery, S. A., & Åsberg, M. (1979). A new depression scale designed to be sensitive to change. *The British Journal of Psychiatry*, *134*(4), 382–389.

Moore II, E., Clements, M. A., Peifer, J. W., & Weisser, L. (2007). Critical analysis of the impact of glottal features in the classification of clinical depression in speech. *IEEE Transactions on Biomedical Engineering*, *55*(1), 96–107.

Morales, M. R., & Levitan, R. (2016). Speech vs. text: A comparative analysis of features for depression detection systems. *Proceedings of the IEEE Spoken Language Technology Workshop*, 136–143.

Mundt, J. C., Snyder, P. J., Cannizzaro, M. S., Chappie, K., & Geralts, D. S. (2007). Voice acoustic measures of depression severity and treatment response collected via interactive voice response (ivr) technology. *Journal of Neurolinguistics*, *20*(1), 50–64.

Mundt, J. C., Vogel, A. P., Feltner, D. E., & Lenderking, W. R. (2012). Vocal acoustic biomarkers of depression severity and treatment response. *Biological Psychiatry*, *72*(7), 580–587.

Muzammel, M., Salam, H., & Othmani, A. (2021). End-to-end multimodal clinical depression recognition using deep neural networks: A comparative analysis. *Computer Methods and Programs in Biomedicine*, *211*, 106433.

Nilsonne, A., & Sundberg, J. (1985). Differences in ability of musicians and nonmusicians to judge emotional state from the fundamental frequency of voice samples. *Music Perception*, 507–516.

Niu, M., Tao, J., Liu, B., & Fan, C. (2019). Automatic depression level detection via lp-norm pooling. *Proceedings of INTERSPEECH*, 4559–4563.

Niu, M., Tao, J., Liu, B., Huang, J., & Lian, Z. (2020). Multimodal spatiotemporal representation for automatic depression level detection. *IEEE Transactions on Affective Computing*.

Nolen-Hoeksema, S., Wisco, B. E., & Lyubomirsky, S. (2008). Rethinking rumination. *Perspectives on Psychological Science*, *3*(5), 400–424.

Nook, E. C., Schleider, J. L., & Somerville, L. H. (2017). A linguistic signature of psychological distancing in emotion regulation. *Journal of Experimental Psychology: General*, *146*(3), 337.

Ozdas, A., Shiavi, R. G., Silverman, S. E., Silverman, M. K., & Wilkes, D. M. (2004). Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk. *IEEE Transactions on Biomedical Engineering*, *51*(9), 1530–1540.

Özseven, T., & Düğenci, M. (2018). Speech acoustic (spac): A novel tool for speech feature extraction and classification. *Applied Acoustics*, *136*, 1–8.

Pan, W., Flint, J., Shenhav, L., Liu, T., Liu, M., Hu, B., & Zhu, T. (2019). Re-examining the robustness of voice features in predicting depression: Compared with baseline of confounders. *PLoS One*, *14*(6), e0218172.

Pan, Z., Gui, C., Zhang, J., Zhu, J., & Cui, D. (2018). Detecting manic state of bipolar disorder based on support vector machine and gaussian mixture model using spontaneous speech. *Psychiatry Investigation*, *15*(7), 695.

Pappagari, R., Wang, T., Villalba, J., Chen, N., & Dehak, N. (2020). x-vectors meet emotions: A study on dependencies between emotion and speaker recognition. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 7169–7173.

Passarotti, A. M., Sweeney, J. A., & Pavuluri, M. N. (2009). Neural correlates of incidental and directed facial emotion processing in adolescents and adults. *Social Cognitive and Affective Neuroscience*, *4*(4), 387–398.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... others (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, *32*.

Patro, S., & Sahu, K. K. (2015). Normalization: A preprocessing stage. *arXiv preprint arXiv:1503.06462*.

Paulhus, D., & Vazire, S. (2007). The self-report method. In R. Robins, R. Fraley, & R. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 224–239). Gilford.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... others (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pentland, A. (2010). *Honest signals: how they shape our world*. MIT Press.

Placidi, G. P., Oquendo, M. A., Malone, K. M., Huang, Y.-Y., Ellis, S. P., & Mann, J. J. (2001). Aggressivity, suicide attempts, and depression: relationship to cerebrospinal fluid monoamine metabolite levels. *Biological Psychiatry*, *50*(10), 783–791.

Quatieri, T. F., & Malyska, N. (2012). Vocal-source biomarkers for depression: A link to psychomotor activity. *Proceedings of INTERSPEECH*.

Qureshi, S. A., Saha, S., Hasanuzzaman, M., & Dias, G. (2019). Multitask representation learning for multimodal estimation of depression level. *IEEE Intelligent Systems*, *34*(5), 45–52.

Rabiner, L., & Schafer, R. (2010). *Theory and applications of digital speech processing*. Prentice Hall Press.

Ravnkilde, B., Videbech, P., Rosenberg, R., Gjedde, A., & Gade, A. (2002). Putative tests of frontal lobe function: a pet-study of brain activation during stroop's test and verbal fluency. *Journal of Clinical and Experimental Neuropsychology*, *24*(4), 534–547.

Ray, A., Kumar, S., Reddy, R., Mukherjee, P., & Garg, R. (2019). Multi-level attention network using text, audio and video for depression prediction. *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 81–88.

Regev, M., Honey, C., Simony, E., & Hasson, U. (2013). Selective and invariant neural responses to spoken and written narratives. *Journal of Neuroscience*, *33*(40), 15978–15988.

Rejaibi, E., Komaty, A., Meriaudeau, F., Agrebi, S., & Othmani, A. (2022). Mfcc-based recurrent neural network for automatic clinical depression recognition and assessment from speech. *Biomedical Signal Processing and Control*, *71*, 103107.

Ringeval, F., Schuller, B., Valstar, M., Cummins, N., Cowie, R., Tavabi, L., ... others (2019). Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition. *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 3–12.

Ringeval, F., Schuller, B., Valstar, M., Gratch, J., Cowie, R., Scherer, S., ... Pantic, M. (2017). Avec 2017: Real-life depression, and affect recognition workshop and challenge. *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, 3–9.

Ritchey, M., Dolcos, F., Eddington, K. M., Strauman, T. J., & Cabeza, R. (2011). Neural correlates of emotional processing in depression: changes with cognitive behavioral therapy and predictors of treatment response. *Journal of Psychiatric Research*, *45*(5), 577–587.

Roberts, G., Lord, A., Frankland, A., Wright, A., Lau, P., Levy, F., ... Breakspear, M. (2017). Functional dysconnection of the inferior frontal gyrus in young people with bipolar disorder or at genetic high risk. *Biological Psychiatry*, *81*(8), 718–727.

Rodrigues Makiuchi, M., Warnita, T., Uto, K., & Shinoda, K. (2019). Multimodal fusion of bert-cnn and gated cnn representations for depression detection. *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 55–63.

Rohanian, M., Hough, J., Purver, M., et al. (2019). Detecting depression with word-level multimodal fusion. *Proceedings of INTERSPEECH*, 1443–1447.

Rolls, E. T., Cheng, W., Gilson, M., Qiu, J., Hu, Z., Ruan, H., ... others (2018). Effective connectivity in depression. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *3*(2), 187–197.

Rubino, I. A., D'Agostino, L., Sarchiola, L., Romeo, D., Siracusano, A., & Docherty, N. M. (2011). Referential failures and affective reactivity of language in schizophrenia and unipolar depression. *Schizophrenia Bulletin*, *37*(3), 554–560.

Rubinstein, R. Y., & Kroese, D. P. (2004). *The cross-entropy method: a unified approach to combinatorial optimization, monte-carlo simulation, and machine learning* (Vol. 133). Springer.

Rude, S., Gortner, E.-M., & 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., ... others (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.

Sahu, S., & Espy-Wilson, C. Y. (2016). Speech features for depression detection. *Proceedings of INTERSPEECH*, 1928–1932.

Saidi, A., Othman, S. B., & Saoud, S. B. (2020). Hybrid cnn-svm classifier for efficient depression detection system. *Proceedings of the 4th International Conference on Advanced Systems and Emergent Technologies*, 229–234.

Santomauro, D. F., Herrera, A. M. M., Shadid, J., Zheng, P., Ashbaugh, C., Pigott, D. M., ... others (2021). Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the covid-19 pandemic. *The Lancet*, *398*(10312), 1700–1712.

Sataloff, R. T., Heman-Ackah, Y. D., & Hawkshaw, M. J. (2007). Clinical anatomy and physiology of the voice. *Otolaryngologic Clinics of North America*, *40*(5), 909–929.

Schaefer, S. M., Jackson, D. C., Davidson, R. J., Aguirre, G. K., Kimberg, D. Y., & Thompson-Schill, S. L. (2002). Modulation of amygdalar activity by the conscious regulation of negative emotion. *Journal of Cognitive Neuroscience*, *14*(6), 913–921.

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

Scherer, K. R. (2013). Vocal assessment of affective disorders: His life and creativity. In *Depression and expressive behavior* (pp. 57–82). Routledge.

Scherer, S., Lucas, G. M., Gratch, J., Rizzo, A. S., & Morency, L.-P. (2015). Self-reported symptoms of depression and ptsd are associated with reduced vowel space in screening interviews. *IEEE Transactions on Affective Computing*, *7*(1), 59–73.

Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT Press.

Schuller, B., & Batliner, A. (2013). *Computational paralinguistics: emotion, affect and personality in speech and language processing*. John Wiley & Sons.

Schuller, B., Batliner, A., Steidl, S., & Seppi, D. (2011). Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, *53*(9-10), 1062–1087.

Schuller, B., Steidl, S., & Batliner, A. (2009). The interspeech 2009 emotion challenge. *Proceedings of INTERSPEECH*.

Scibelli, F., Roffo, G., Tayarani, M., Bartoli, L., De Mattia, G., Esposito, A., & Vinciarelli, A. (2018). Depression speaks: Automatic discrimination between depressed and non-depressed speakers based on nonverbal speech features. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 6842–6846.

Seghier, M. L., Lazeyras, F., Pegna, A. J., Annoni, J.-M., Zimine, I., Mayer, E., ... Khateb, A. (2004). Variability of fmri activation during a phonological and semantic language task in healthy subjects. *Human Brain Mapping*, *23*(3), 140–155.

Seshadri, G., & Yegnanarayana, B. (2009). Perceived loudness of speech based on the characteristics of glottal excitation source. *The Journal of the Acoustical Society of America*, *126*(4), 2061–2071.

Sethu, V., Ambikairajah, E., & Epps, J. (2009). Speaker dependency of spectral features and speech production cues for automatic emotion classification. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 4693–4696.

Shafritz, K. M., Collins, S. H., & Blumberg, H. P. (2006). The interaction of emotional and cognitive neural systems in emotionally guided response inhibition. *Neuroimage*, *31*(1), 468–475.

Shannon, T. T. E., Lan, S. S., et al. (2016). Speech analysis and depression. *Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 1–4.

Sheline, Y. I., Price, J. L., Yan, Z., & Mintun, M. A. (2010). Resting-state functional mri in depression unmasks increased connectivity between networks via the dorsal nexus. *Proceedings of the National Academy of Sciences*, *107*(24), 11020–11025.

Shen, J., Zhang, X., Huang, X., Wu, M., Gao, J., Lu, D., ... Hu, B. (2020). An optimal channel selection for eeg-based depression detection via kernel-target alignment. *IEEE Journal of Biomedical and Health Informatics*, *25*(7), 2545–2556.

Shen, Y., Yang, H., & Lin, L. (2022). Automatic depression detection: An emotional audio-textual corpus and a gru/bilstm-based model. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 6247–6251.

Shrout, P. E., Stadler, G., Lane, S. P., McClure, M. J., Jackson, G. L., Clavél, F. D., ... Bolger, N. (2018). Initial elevation bias in subjective reports. *Proceedings of the National Academy of Sciences*, *115*(1), E15–E23.

Siegle, G. J., Steinhauer, S. R., Thase, M. E., Stenger, V. A., & Carter, C. S. (2002). Can't shake that feeling: event-related fmri assessment of sustained amygdala activity in response to emotional information in depressed individuals. *Biological Psychiatry*, *51*(9), 693–707.

Silva, W. J., Lopes, L., Galdino, M. K. C., & Almeida, A. A. (2021). Voice acoustic parameters as predictors of depression. *Journal of Voice*.

Simantiraki, O., Charonyktakis, P., Pampouchidou, A., Tsiknakis, M., & Cooke, M. (2017). Glottal source features for automatic speech-based depression assessment. *Proceedings of INTERSPEECH*, 2700–2704.

Singh, R. (2019). *Profiling humans from their voice* (Vol. 41). Springer.

Snyder, D., Garcia-Romero, D., Povey, D., & Khudanpur, S. (2017). Deep neural network embeddings for text-independent speaker verification. *Proceedings of INTERSPEECH*, *2017*, 999–1003.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018). X-vectors: Robust dnn embeddings for speaker recognition. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 5329–5333.

Song, S., Jaiswal, S., Shen, L., & Valstar, M. (2020). Spectral representation of behaviour primitives for depression analysis. *IEEE Transactions on Affective Computing*.

Stanek, M., & Polak, L. (2013). Algorithms for vowel recognition in fluent speech based on formant positions. *Proceedings of the 36th International Conference on Telecommunications and Signal Processing*, 521–525.

Stasak, B., & Epps, J. (2017). Differential performance of automatic speech-based depression classification across smartphones. *Proceedings of the 7th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos*, 171–175.

Stasak, B., Huang, Z., Epps, J., & Joachim, D. (2021). Depression classification using n-gram speech errors from manual and automatic stroop color test transcripts. *Proceedings of the 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society*, 1631–1635.

Stasak, B., Joachim, D., & Epps, J. (2022). Breaking age barriers with automatic voice-based depression detection. *IEEE Pervasive Computing*, *21*(2), 10–19.

Stolar, M. N., Lech, M., Stolar, S. J., & Allen, N. B. (2018). Detection of adolescent depression from speech using optimised spectral roll-off parameters. *Biomedical Journal*, *2*, 10.

Sturim, D., Torres-Carrasquillo, P. A., Quatieri, T. F., Malyska, N., & McCree, A. (2011). Automatic detection of depression in speech using gaussian mixture modeling with factor analysis. *Proceedings of INTERSPEECH*.

Sun, H., Wang, H., Liu, J., Chen, Y.-W., & Lin, L. (2022). Cubemlp: An mlp-based model for multimodal sentiment analysis and depression estimation. *Proceedings of the 30th ACM International Conference on Multimedia*, 3722–3729.

Sundberg, J., Patel, S., Bjorkner, E., & Scherer, K. R. (2011). Interdependencies among voice source parameters in emotional speech. *IEEE Transactions on Affective Computing*, *2*(3), 162–174.

Tackman, A. M., Sbarra, D. A., Carey, A. L., Donnellan, M. B., Horn, A. B., Holtzman, N. S., ... Mehl, M. R. (2019). Depression, negative emotionality, and self-referential language: A multi-lab, multi-measure, and multi-language-task research synthesis. *Journal of Personality and Social Psychology*, *116*(5), 817.

Taguchi, T., Tachikawa, H., Nemoto, K., Suzuki, M., Nagano, T., Tachibana, R., ... Arai, T. (2018). Major depressive disorder discrimination using vocal acoustic features. *Journal of Affective Disorders*, *225*, 214–220.

Tanner, J.-A., Hensel, J., Davies, P. E., Brown, L. C., Dechairo, B. M., & Mulsant, B. H. (2020). Economic burden of depression and associated resource use in manitoba, canada. *The Canadian Journal of Psychiatry*, *65*(5), 338–346.

Tao, F., Esposito, A., & Vinciarelli, A. (2020). Spotting the traces of depression in read speech: An approach based on computational paralinguistics and social signal processing. *Proceedings of INTERSPEECH*, 1828–1832.

Tao, F., Esposito, A., & Vinciarelli, A. (2023). The androids corpus: A new publicly available benchmark for speech based depression detection. *Proceedings of INTERSPEECH*, to be published.

Tao, F., Ge, X., Ma, W., Esposito, A., & Vinciarelli, A. (2023). Multi-local attention for speech-based depression detection. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1–5.

Tao, F., Ma, W., Ge, X., Esposito, A., & Vinciarelli, A. (2023). The relationship between speech features changes when you get depressed: Feature correlations for improving speed and performance of depression detection. *arXiv preprint arXiv:2307.02892*.

Tasnim, M., Ehghaghi, M., Diep, B., & Novikova, J. (2022). Depac: a corpus for depression and anxiety detection from speech. *Proceedings of the 8th Workshop on Computational Linguistics and Clinical Psychology*, 1–16.

Tayarani, M., Esposito, A., & Vinciarelli, A. (2019). What an" ehm" leaks about you: Mapping fillers into personality traits with quantum evolutionary feature selection algorithms. *IEEE Transactions on Affective Computing*.

Teixeira, J. P., Oliveira, C., & Lopes, C. (2013). Vocal acoustic analysis–jitter, shimmer and hnr parameters. *Procedia Technology*, *9*, 1112–1122.

Tieleman, T., Hinton, G., et al. (2012). Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, *4*(2), 26–31.

Tølbøll, K. B. (2019). Linguistic features in depression: a meta-analysis. *Journal of Language Works-Sprogvidenskabeligt Studentertidsskrift*, *4*(2), 39–59.

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

Tsao, C.-W., Lin, Y.-S., Chen, C.-C., Bai, C.-H., & Wu, S.-R. (2006). Cytokines and serotonin transporter in patients with major depression. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *30*(5), 899–905.

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

Valstar, M., Schuller, B., Smith, K., Almaev, T., Eyben, F., Krajewski, J., ... Pantic, M. (2014). Avec 2014: 3d dimensional affect and depression recognition challenge. *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, 3–10.

Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., ... Pantic, M. (2013). Avec 2013: the continuous audio/visual emotion and depression recognition challenge. *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, 3–10.

Van der Zanden, R., Curie, K., Van Londen, M., Kramer, J., Steen, G., & Cuijpers, P. (2014). Web-based depression treatment: Associations of clients' word use with adherence and outcome. *Journal of Affective Disorders*, *160*, 10–13.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Vicsi, K., Sztahó, D., & Kiss, G. (2012). Examination of the sensitivity of acoustic-phonetic parameters of speech to depression. *Proceedings of the IEEE 3rd International Conference on Cognitive Infocommunications*, 511–515.

Victor, T. A., Furey, M. L., Fromm, S. J., Öhman, A., & Drevets, W. C. (2010). Relationship between amygdala responses to masked faces and mood state and treatment in major depressive disorder. *Archives of General Psychiatry*, *67*(11), 1128–1138.

Vinciarelli, A., Pantic, M., & Bourlard, H. (2009). Social Signal Processing: Survey of an emerging domain. *Image and Vision Computing*, *27*(12), 1743–1759.

Vos, T., Allen, C., Arora, M., Barber, R. M., Bhutta, Z. A., Brown, A., ... others (2016). Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the global burden of disease study 2015. *The Lancet*, *388*(10053), 1545–1602.

Wager, T. D., Davidson, M. L., Hughes, B. L., Lindquist, M. A., & Ochsner, K. N. (2008). Prefrontal-subcortical pathways mediating successful emotion regulation. *Neuron*, *59*(6), 1037–1050.

Wagner, S., Doering, B., Helmreich, I., Lieb, K., & Tadić, A. (2012). A meta-analysis of executive dysfunctions in unipolar major depressive disorder without psychotic symptoms and their changes during antidepressant treatment. *Acta Psychiatrica Scandinavica*, *125*(4), 281–292.

Walfish, S., McAlister, B., O'Donnell, P., & Lambert, M. J. (2012). An investigation of self-assessment bias in mental health providers. *Psychological Reports*, *110*(2), 639–644.

Wanderley Espinola, C., Gomes, J. C., Mônica Silva Pereira, J., & dos Santos, W. P. (2022). Detection of major depressive disorder, bipolar disorder, schizophrenia and generalized anxiety disorder using vocal acoustic analysis and machine learning: an exploratory study. *Research on Biomedical Engineering*, *38*(3), 813–829.

Wang, J., Zhang, L., Liu, T., Pan, W., Hu, B., & Zhu, T. (2019). Acoustic differences between healthy and depressed people: a cross-situation study. *BMC Psychiatry*, *19*, 1–12.

Wang, Q., & Liu, N. (2022). Speech detection of depression based on multi-mlp. *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine*, 3896–3898.

Wang, R., Hao, Y., Yu, Q., Chen, M., Humar, I., & Fortino, G. (2021). Depression analysis and recognition based on functional near-infrared spectroscopy. *IEEE Journal of Biomedical and Health Informatics*, *25*(12), 4289–4299.

Wang, Y., Lu, X., & Shi, D. (2022). Mfcc-based deep convolutional neural network for audio depression recognition. *Proceedings of the International Conference on Asian Language Processing*, 162–166.

Watson, D., Clark, L. A., & Carey, G. (1988). Positive and negative affectivity and their relation to anxiety and depressive disorders. *Journal of Abnormal Psychology*, *97*(3), 346.

WHO, W. (2017). Depression and other common mental disorders: global health estimates. *Depress Other Common Ment Disord Glob Heal Estim*.

Williams, S. Z., Chung, G. S., & Muennig, P. A. (2017). Undiagnosed depression: A community diagnosis. *SSM-population Health*, *3*, 633–638.

Williamson, J. R., Bliss, D. W., & Browne, D. W. (2011). Epileptic seizure prediction using the spatiotemporal correlation structure of intracranial eeg. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 665–668.

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

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

Williamson, J. R., Quatieri, T. F., Helfer, B. S., Horwitz, R., Yu, B., & Mehta, D. D. (2013). Vocal biomarkers of depression based on motor incoordination. *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, 41–48.

Williamson, J. R., Young, D., Nierenberg, A. A., Niemi, J., Helfer, B. S., & Quatieri, T. F. (2019). Tracking depression severity from audio and video based on speech articulatory coordination. *Computer Speech & Language*, *55*, 40–56.

Wise, T., Marwood, L., Perkins, A., Herane-Vives, A., Joules, R., Lythgoe, D., … Cleare, A. (2017). Instability of default mode network connectivity in major depression: a two-sample confirmation study. *Translational Psychiatry*, *7*(4), e1105–e1105.

Wu, P., Wang, R., Lin, H., Zhang, F., Tu, J., & Sun, M. (2022). Automatic depression recognition by intelligent speech signal processing: A systematic survey. *CAAI Transactions on Intelligence Technology*.

Xie, Z., Zinszer, B. D., Riggs, M., Beevers, C. G., & Chandrasekaran, B. (2019). Impact of depression on speech perception in noise. *PloS One*, *14*(8), e0220928.

Xu, K., Hu, W., Leskovec, J., & Jegelka, S. (2018). How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*.

Xu, X., Chikersal, P., Dutcher, J. M., Sefidgar, Y. S., Seo, W., Tumminia, M. J., … others (2021). Leveraging collaborative-filtering for personalized behavior modeling: a case study of depression detection among college students. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *5*(1), 1–27.

Xu, X., Peng, H., Bhuiyan, M. Z. A., Hao, Z., Liu, L., Sun, L., & He, L. (2021). Privacy-preserving federated depression detection from multisource mobile health data. *IEEE Transactions on Industrial Informatics*, *18*(7), 4788–4797.

Yamamoto, M., Takamiya, A., Sawada, K., Yoshimura, M., Kitazawa, M., Liang, K.-c., … Kishimoto, T. (2020). Using speech recognition technology to investigate the association between timing-related speech features and depression severity. *PloS One*, *15*(9), e0238726.

Yang, H., Cusin, C., & Fava, M. (2005). Is there a placebo problem in antidepressant trials? *Current Topics in Medicinal Chemistry*, *5*(11), 1077–1086.

Yang, L., Jiang, D., & Sahli, H. (2021). Integrating deep and shallow models for multi-modal depression analysis—hybrid architectures. *IEEE Transactions on Affective Computing*, *12*(1), 239-253.

Yang, Y., Fairbairn, C., & Cohn, J. F. (2012). Detecting depression severity from vocal prosody. *IEEE Transactions on Affective Computing*, *4*(2), 142–150.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical attention networks for document classification. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1480–1489.

Yeh, C.-F., Mahadeokar, J., Kalgaonkar, K., Wang, Y., Le, D., Jain, M., ... Seltzer, M. L. (2019). Transformer-transducer: End-to-end speech recognition with self-attention. *arXiv preprint arXiv:1910.12977*.

Yin, F., Du, J., Xu, X., & Zhao, L. (2023). Depression detection in speech using transformer and parallel convolutional neural networks. *Electronics*, *12*(2), 328.

Yingthawornsuk, T., Keskinpala, H. K., France, D., Wilkes, D. M., Shiavi, R. G., & Salomon, R. M. (2006). Objective estimation of suicidal risk using vocal output characteristics. *Proceedings of the 9th International Conference on Spoken Language Processing*.

Young, R. C., Biggs, J. T., Ziegler, V. E., & Meyer, D. A. (1978). A rating scale for mania: reliability, validity and sensitivity. *The British Journal of Psychiatry*, *133*(5), 429–435.

Yue, C., Ware, S., Morillo, R., Lu, J., Shang, C., Bi, J., ... Wang, B. (2020). Automatic depression prediction using internet traffic characteristics on smartphones. *Smart Health*, *18*, 100137.

Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2007). A survey of affect recognition methods: audio, visual and spontaneous expressions. *Proceedings of the 9th International Conference on Multimodal Interfaces*, 126–133.

Zhao, Y., Liang, Z., Du, J., Zhang, L., Liu, C., & Zhao, L. (2021). Multi-head attention-based long short-term memory for depression detection from speech. *Frontiers in Neurorobotics*, *15*, 684037.

Zhao, Y., Xie, Y., Liang, R., Zhang, L., Zhao, L., & Liu, C. (2021). Detecting depression from speech through an attentive lstm network. *IEICE TRANSACTIONS on Information and Systems*, *104*(11), 2019–2023.

Zhao, Z., Bao, Z., Zhang, Z., Cummins, N., Wang, H., & Schuller, B. (2020). Hierarchical attention transfer networks for depression assessment from speech. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 7159–7163.

Zhao, Z., Bao, Z., Zhang, Z., Deng, J., Cummins, N., Wang, H., ... Schuller, B. (2019). Automatic assessment of depression from speech via a hierarchical attention transfer network and attention autoencoders. *IEEE Journal of Selected Topics in Signal Processing*, *14*(2), 423–434.

Zhu, Y., & Jiang, S. (2019). Attention-based densely connected lstm for video captioning. *Proceedings of the 27th ACM International Conference on Multimedia*, 802–810.

Zhu, Y., Kim, Y.-C., Proctor, M. I., Narayanan, S. S., & Nayak, K. S. (2012). Dynamic 3-d visualization of vocal tract shaping during speech. *IEEE Transactions on Medical Imaging*, *32*(5), 838–848.

Zimmermann, J., Wolf, M., Bock, A., Peham, D., & Benecke, C. (2013). The way we refer to ourselves reflects how we relate to others: Associations between first-person pronoun use and interpersonal problems. *Journal of Research in Personality*, *47*(3), 218–225.