The Rhythm of Depressed Speech: An Analysis of Timing Variabilities

Senior Thesis

Presented to
The Faculty of the School of Arts and Sciences
Brandeis University

Undergraduate Program in
Language and Linguistics
Keith Plaster, advisor

In partial fulfillment of the requirements
for the degree of
Bachelor of Arts

by
Adrianne Schwarzentuber
May 2016
# Contents

0 Introduction .............................. 1

1 Literature Review .......................... 5
   1.1 Automatic Evaluation of Biomarkers .... 6
   1.2 Major Depressive Disorder ............ 7
   1.3 Vocal Features in Depression ......... 9
   1.4 Speech Rhythm ......................... 13
      1.4.1 Early Theory ...................... 14
      1.4.2 Disproving Isochrony ............ 16
      1.4.3 Phonotactic Differences ........... 17
      1.4.4 Rhythm Metrics ................... 18
      1.4.5 Applications of the Rhythm Metrics . 23
      1.4.6 Criticisms of the Metrics ....... 25
   1.5 Depression, Dysarthria & Speech Rhythm . 27

2 Methodology ............................... 30
   2.1 Speech Rhythm Metrics ................. 30
   2.2 Data .................................... 33
   2.3 Automatic Speech Segmentation ....... 36
   2.4 Calculation of Metrics and Statistics ... 37

3 Results .................................. 39
   3.1 Group Analysis: All Subjects ........... 41
      3.1.1 Free Speech ....................... 41
      3.1.2 Alphabet .......................... 42
      3.1.3 Grandfather Passage ............... 43
   3.2 Group Analysis: Females ................. 44
      3.2.1 Free Speech ....................... 45
      3.2.2 Alphabet .......................... 45


List of Figures

1.1 A visual representation of different isochronous units in syllable-timed and stress-timed languages. ................................................. 15
1.2 An example of the stress-based continuum, taken from Dauer (1983). ............... 18
1.3 The process of segmenting the phrase “great ideas” into consonant and vowel intervals. 20
1.4 The plot of %V by ΔC from Ramus et al. (1999), demonstrating a separation of languages into their rhythm classes. ................................. 21
1.5 The process of segmenting the phrase “great ideas” into CV intervals. .................. 23
3.1 Significant correlations in the alphabet task ............................................. 44
3.2 Significant correlations in the Grandfather Passage task. ................................. 44
3.3 Significant correlations in the alphabet task in the female subgroup. .................. 46
3.4 Significant correlations in the Grandfather Passage task in the female subgroup. ...... 47
3.5 Significant correlations in the alphabet task in the male subgroup. ..................... 49
3.6 Significant correlations in the Grandfather Passage task in the male subgroup. ....... 49
3.7 Significant correlations in the Grandfather Passage task in the older subgroup. ...... 52
3.8 Significant correlations in the alphabet task in the younger subgroup. ................ 54
3.9 Significant correlations in the Grandfather Passage task in the younger subgroup. ... 54
4.1 All significant correlations for each metric across all data. ................................. 58
### List of Tables

1.1 A summary of features found to be significant in studies of depressed speech . . . . 11
1.2 A list of Dauer’s rhythm features for the classification of languages (Dauer 1987) . . 19
2.1 A list of the rhythm and other metrics, with their definitions, calculated in this analysis. . . 32
2.2 p and r values for Spearman correlations calculated between depression scales. . . . 35
3.1 Results from the Wilcoxon Mann-Whitney tests between the metrics and gender. . . 40
3.2 All significant (p<.05) correlations between age and the metric values. . . . . 41
3.3 The number of samples in each task across all participants. . . . . . . . . 41
3.4 The number of samples in each task across in the female subgroup. . . . . . . . 44
3.5 The number of samples in each task in the male subgroup. . . . . . . . . 47
3.6 The number of samples in each task in the older subgroup. . . . . . . . . 50
3.7 The number of samples in each task in the younger subgroup. . . . . . . . 52
3.8 A summary of all correlation results. . . . . . . . . . . . . . . . . . . . . . . 55
I would like to express my sincerest gratitude to my advisor, Professor Keith Plaster. He has spent countless hours discussing, reading, and providing feedback on my work, and this thesis would not exist without his guidance. Thank you for being excited about my work when I was not, and for being a source of constant encouragement.

I also extend thanks to Group 48 of the MIT Lincoln Laboratory for providing me with this research opportunity.* Specifically, thanks go to Tom Quatieri for supervising this project, Elizabeth Godoy for all of the many ways she has helped me to analyze, interpret, and present my data over the past year, Jennifer Melot for serving as my Hydra liaison, and to Adam Lammert for teaching me how to use MATLAB and contributing his ideas.

I’d like to thank my parents for remaining my biggest fans, despite the fact that I’m not going to med school. Finally, thank you, David, for enduring long hours as I work, and for providing endless support and chocolate.

*This work is sponsored by the Assistant Secretary of Defense for Research & Engineering under Air Force contract #FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.
Introduction

Depression is a widely prevalent disease in the U.S., with 7.6% of Americans experiencing a depressive episode each year (Pratt & Brody 2014). Current diagnosis and monitoring methods tend to rely heavily on subjective clinician observations and patient self-report mood scales, and accurate diagnosis and monitoring require consultation with a trained clinician, which can quickly become expensive. Given these issues, the creation of an objective method of depression evaluation that is administrable outside of the office of a medical professional would be helpful for the management of the disorder.

There is currently much focus on the development of automatic tools for the detection of disorders, including depression, which would help mitigate the subjectivity inherent in human obser-
vations and self-reflections. By identifying physical signals not perceptible to humans, these tools could allow for earlier detection of depression, while also reducing the frequency of clinician visits to monitor progression of the depression. Ultimately, such tools would increase clinicians’ ability to detect, monitor, and treat depression.

In particular, an ideal candidate for the basis of automatic tools is speech, a complex neurocognitive and motor process. Variations in speech production associated with physiological changes, which are referred to as vocal biomarkers, are founded in the idea that changes in brain function will manifest as changes in motor output. Speech analysis is an excellent evaluation tool because data collection is non-invasive, relatively straightforward, and inexpensive.

Several studies have previously explored variations in speech of depressed patients and help motivate the current analysis. It has been shown that pause length (Ellgring & Scherer 1996; Mundt et al. 2007) and frequency (Mundt et al. 2007) increase with increased depression, demonstrating that more depressed speech contains more overall silence. Speech rate (Ellgring & Scherer 1996; Mundt et al. 2007; Cannizzaro et al. 2004; Horwitz et al. 2013) and phone rate (Trevino et al. 2011), which measures the frequency of individual segments rather than words or syllables, have been shown to decrease with higher depression severity, indicating that more depressed patients speak more slowly. Other acoustic features have also been investigated. For example, pitch variability (Mundt et al. 2007; Cannizzaro et al. 2004) has been shown to decrease with depression, suggesting that depressed patients speak in more monotone than healthy counterparts. Many of these features can be attributed to the effects of psychomotor retardation, a common occurrence in depression which leads to the slowing of articulator movements (Buyukdura et al. 2011).

As in depression, psychomotor retardation also occurs in Parkinson’s disease. Parkinsonian dysarthria is a motor speech disorder that occurs as a result of injury to the basal ganglia and involves disrupted articulation. It has been characterized by various speech features that have also been found to be relevant in depression. Given that psychomotor retardation occurs in both, and that both dis-
states have an underlying impact on a common area of the brain, specifically the basal ganglia (Duffy 2013; Sobin & Sackeim 1997), it is reasonable to posit that both depression and Parkinson’s may have similar effects on speech.

With this underlying connection in mind, we may look to speech features that have been investigated in dysarthria but not yet in depression, and speech rhythm is such a feature. Altered or disrupted rhythm is often cited as a symptom of dysarthria, and Liss et al. (1998) found success using speech rhythm metrics to distinguish between normal and dysarthric speech. Speech rhythm metrics capture relative timing variabilities on a consonant-vowel (CV) level, while some measurements also encapsulate speech rate. As various articulation rate measurements have previously been shown to be significant in depression (Ellgring & Scherer 1996; Mundt et al. 2007; Cannizzaro et al. 2004; Horwitz et al. 2013; Trevino et al. 2011), and due to its measurement of local timing variabilities and sensitivity to speech rate, speech rhythm is a possible informative feature. Consequently, this analysis explores whether speech rhythm features co-vary with depression in a way such that they could potentially be used for future classification and monitoring of depression.

To investigate the changes in speech rhythm that occur in depression, I analyzed rhythm metrics in a collection of speech recordings from Mundt et al. (2007) made by depressed patients as they underwent a six-week treatment period. The rhythm metrics were calculated via a Python script designed to take the output of automatic speech segmentation to identify consonant and vowel intervals, as described in Ramus et al. (1999), and which calculates and outputs the rhythm metric values. To determine the amount of co-variation between the metric values and depression severity, I calculated correlations between the rhythm metric values and depression severity scores, which were included in the Mundt et al. (2007) data. In addition to analyzing speech tasks individually, I further divided the data into subgroups based on gender and age to explore potential differences in rhythm in these subgroups.

This thesis is organized as follows: Chapter 1 discusses the existing literature on biomarkers, de-
CHAPTER 0. INTRODUCTION

pression, speech features in depression, and speech rhythm, and further explains a parallel between depression and dysarthria. Chapter 2 describes the methodologies used in the analysis, specifically the speech rhythm metrics, the speech data, and speech processing techniques. Chapter 3 presents the results of the analyses, and Chapter 4 interprets and discusses those results. Finally, Chapter 5 gives a brief summary of the conclusions of the thesis and identifies directions for future research.
This chapter reviews previous and current literature in multiple fields related to the topic of this thesis. It begins with an explanation of biomarkers and summarizes some of their current applications. Then, it moves into a brief review of depression, including prominent symptoms and potential causes. Following the section on depression is a review of various studies that have explored speech features in depression. Next, I discuss the history of the study of speech rhythm, moving from early theory to more current thought and methods, including a discussion of speech rhythm metrics. Finally, the literature review outlines a connection between depressed and dysarthric speech and provides motivation for investigating speech rhythm in depression.
1.1 Automatic Evaluation of Biomarkers

Of growing interest and importance to the medical and psychological communities is the existence of biological markers (“biomarkers”) and their potential to inform automatic evaluation and monitoring tools. Biomarkers are based in the idea that pathogens, pharmaceuticals, and normal body processes produce measurable responses that can be used to assess physiological functioning and reactions. Familiar biomarkers include blood pressure, temperature, cholesterol level, and other physiological measurements, which provide information about the health or state of a person. Current research focuses on the identification of biomarkers that can indicate the presence of particular states or disorders, toward the goal of developing automatic diagnostic and monitoring tools.

Speech and facial expression are the subjects of present investigations in both state detection and disorder detection (e.g. López-de Ipiña et al. (2013); Kiss et al. (2012); Vanello et al. (2012); Ringeval et al. (2012)). Both processes require complex coordination of articulatory muscles that must be precisely timed and implemented by neuromotor circuits. Any disruption to motor-related areas of the brain due to cognitive, emotive, or pathogenic conditions may be reflected in alterations to or disruptions in the motor output. Therefore, by quantifying and characterizing these changes, researchers aim to quantify and characterize the underlying variations occurring in the brain. This thesis will focus specifically on speech and the ways that it can reveal information about brain state. Speech is an ideal biomarker because it is inexpensive, uncomplicated to collect, and non-invasive to the patient. Speech features have seen initial success in multiple studies and hold promise as informative biomarkers.

Vocal biomarkers have been shown to be relevant in the automatic detection of a number of states, including emotion, fatigue, and cognitive load. Often, features like $F_0$ mean and range, intensity, and speaking rate vary in ways that can indicate the presence of a particular state. In emotion, Ringeval et al. (2012) successfully utilized speech features for multimodal emotion detection.
CHAPTER 1. LITERATURE REVIEW 1.2. MAJOR DEPRESSIVE DISORDER

as part of an Audio/Visual Emotion Challenge. According to Kreiman & Sidtis (2011), expression of emotion heavily involves the basal ganglia, which are primarily involved in motor function. This intimate connection between speech and emotion explains why speech can be used in emotion detection. With respect to fatigue, Vogel et al. (2010) determined that as fatigue increased, speaking rate decreased, mean pause time increased, and total speech time increased. Finally, regarding cognitive load, a study in healthy adults by Cohen et al. (2015) determined that speech production is affected by information processing load, with larger loads producing longer pauses and less variability in frequency and intensity levels.

Vocal biomarkers are also being used to determine the existence and severity of disorders, including psychiatric disorders and neurodegenerative diseases that compromise motor functioning. For example, Vanello et al. (2012) used pitch and jitter features to find differences between mood states in bipolar patients, while Novotny et al. (2014) used vocal and articulatory features, like voice quality, laryngeal coordination, and tongue movement, to distinguish speech of Parkinson’s patients and healthy controls, as did Laaridh et al. (2015). Skodda & Schlegel (2008) found that Parkinson’s patients exhibited altered speech features as compared to controls. Additionally using automatically extracted speech features, López-de Ipiña et al. (2013) found promising results in the detection of early Alzheimer’s disease.

One area in which biomarkers could be particularly useful is in Major Depressive Disorder (MDD), which can be difficult to both diagnose and monitor. The burden of MDD continues to increase, and we will now turn our attention to developing vocal biomarkers and automatic tools for its detection and diagnosis.

1.2 MAJOR DEPRESSIVE DISORDER

Depression, clinically classified as Major Depressive Disorder (see Appendix C.1 for diagnostic criteria), is one of the most prevalent mood disorders in the United States, affecting 7.6% of the pop-
population annually (Pratt & Brody 2014), occurring 1.5-3 times more often in females (American Psychiatric Association 2013), and most severely affecting women aged 40-59 (Pratt & Brody 2014). In addition to its immense emotional and physical toll, a 2013 study estimated that absenteeism due to depression costs U.S. employers an annual $23 billion (Witters et al. 2013). Its massive pervasiveness and elusive etiology has prompted much research into its causes and effective treatments.

Depression is a mood disorder characterized by depressed mood and loss of interest or pleasure in previously enjoyed activities (American Psychiatric Association 2013). Although the exact causes of depression are yet unknown, some researchers believe that it occurs because of neurological changes such as disruption in neurotransmitter uptake and regulation and hippocampus abnormalities (Kramer 2005; Mondimore 2006). Others maintain that genetic predisposition and irregular or disturbed gene expression contribute to an extensive cascade of physiological changes that make it difficult to determine the underlying cause (Schweitzer & Tuckwell 2004).

A common symptom of depression is psychomotor retardation, which is listed as one of its diagnostic criteria (American Psychiatric Association 2013). Psychomotor retardation is one of the more easily observed physiological effects of depression, often typified by slowing of speech, facial expressions, and other motor movements. With regards to speech specifically, disturbances frequently affect pause times, volume, tone, and prosody. In some instances, the severity of psychomotor retardation has been found to be correlated with depression severity (Buyukdura et al. 2011).

However, psychomotor retardation is not the only cause for speech changes associated with depression; Christopher & MacDonald (2005) discovered that depression also results in cognitive effects, such as disrupting working memory. In the speech production process, working memory is thought to be used to store phonetic and prosodic information during speech planning and perception, so disruptions to working memory also produce disruptions in speech (Baddeley 2003).

Thus, given the common occurrence of motor retardation1 and cognitive disruptions in depres-

1Psychomotor retardation does not occur in every instance of MDD. However, it occurs frequently
sion, observable and quantifiable changes in the speech output of depressed patients as compared to non-depressed persons are expected.

1.3 Vocal Features in Depression

In fact, many previous studies have examined depressed speech in search of acoustic or prosodic features that could serve as indicators of the existence and severity of depression. Most often, these features are acoustically based. The following section will explore these studies and discuss features found to be significant in each. For a summary of these studies and their significant results, see Table 1.1.

A pilot study by Darby et al. (1984) found reduced use of stress patterns, monopitch, monoloudness, reduced pitch, reduced loudness, harsh quality, and loudness decay to be exhibited abnormally in depressed patients as compared to controls. In the study, depressed and control male subjects completed voice recordings (including reading the Grandfather Passage, holding a conversation, and a phoneme repetition task) over the course of six weeks. Trained raters blindly rated each of the recordings on 40 dimensions, each on a seven-point scale. Dimensions included categories like monopitch, intelligibility, and hypernasality, and points were added up to determine a total speech score. They found that the pretreatment speech scores for depressed patients were significantly higher than those of controls, indicating that the depressed speech was perceived to be different from the control speech. Additionally, a majority of depressed patients exhibited abnormalities in the previously mentioned seven particular dimensions. However, the authors noted that the depressed speech did not sound strikingly different from that of the control subjects.

Ellgring & Scherer (1996) similarly analyzed depressed speech samples and found that speech rate enough to be regarded as a diagnostic symptom.

2This feature is not defined in Darby et al. (1984). However, it likely refers to reduced differentiation between stressed and unstressed syllables.

3Lacking in pitch variation.

4Lacking in loudness variation.
was slower and mean pause duration was longer when patients were in a depressed state. They obtained speech samples from standardized clinical interviews of hospitalized depressed patients over the course of their treatment. By comparing the depressed and recovered stages, they determined that as depression level decreased, speech rate, as measured by the frequency of syllables, increased, and relatedly, mean pause duration decreased, indicating that patients in a depressed state spoke more slowly and produced longer pauses. For women only, it was found that $F_0$ minimum decreased and $F_0$ range increased significantly in a recovered state. Thus, in a depressed state, female patients spoke with a higher and less variable pitch. The authors state that though some of the results are explained by psychomotor retardation, the gender differences are not, and suggest that further research into gender differences, speech, and depression should be considered. These results also mirror those found by Vogel et al. (2010) in a study on fatigue, which also often occurs in depression (American Psychiatric Association 2013).

Using clinician-rated scores of videotaped interviews, Cannizzaro et al. (2004) found that with increased depression severity, speaking rate decreased, as did pitch variation (though not significantly), while pause time moderately increased. These results suggest slower speech, lengthier pauses, and less variable pitch in a depressed state, echoing the findings of Ellgring & Scherer (1996).

In a longitudinal study, Mundt et al. (2007) additionally found differences in pitch variability, speaking rate, and silence length between depressed and non-depressed states. They collected speech samples of depressed patients as they underwent treatment for depression. Analyses revealed that higher depression severity was correlated with longer utterance durations due to longer and more variable pauses. The depressed speech of participants that responded to treatment exhibited less $F_0$ variability, longer and more frequent pause time, and decreased speaking rate in their depressed states, replicating previous findings.

The MIT Lincoln Laboratory has completed multiple analyses using the Mundt et al. (2007)
### Table 1.1: A summary of features found to be significant in studies of depressed speech

<table>
<thead>
<tr>
<th>Study</th>
<th>Significant Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Darby et al. (1984)</td>
<td>Reduced use of stress patterns</td>
</tr>
<tr>
<td></td>
<td>Monopitch</td>
</tr>
<tr>
<td></td>
<td>Monoloudness</td>
</tr>
<tr>
<td></td>
<td>Reduced pitch</td>
</tr>
<tr>
<td></td>
<td>Reduced loudness</td>
</tr>
<tr>
<td></td>
<td>Harsh quality</td>
</tr>
<tr>
<td></td>
<td>Loudness decay</td>
</tr>
<tr>
<td>Ellgring &amp; Scherer (1996)</td>
<td>Speech (syllable) rate</td>
</tr>
<tr>
<td></td>
<td>Mean pause duration</td>
</tr>
<tr>
<td></td>
<td>$F_0$ minimum (for women only)</td>
</tr>
<tr>
<td></td>
<td>$F_0$ range (for women only)</td>
</tr>
<tr>
<td>Cannizzaro et al. (2004)</td>
<td>$F_0$ variability</td>
</tr>
<tr>
<td></td>
<td>Pause time</td>
</tr>
<tr>
<td></td>
<td>Speech rate</td>
</tr>
<tr>
<td>Mundt et al. (2007)</td>
<td>$F_0$ variability</td>
</tr>
<tr>
<td></td>
<td>Pause time</td>
</tr>
<tr>
<td></td>
<td>Speech rate</td>
</tr>
<tr>
<td>Trevino et al. (2011)</td>
<td>Speech (phone) rate</td>
</tr>
<tr>
<td>Quatieri &amp; Malyska (2012)</td>
<td>Shimmer</td>
</tr>
<tr>
<td></td>
<td>Aspiration</td>
</tr>
<tr>
<td>Horwitz et al. (2013)</td>
<td>Shimmer</td>
</tr>
<tr>
<td></td>
<td>Jitter</td>
</tr>
<tr>
<td></td>
<td>Speech rate</td>
</tr>
</tbody>
</table>

data. A study by Quatieri & Malyska (2012) revealed that as depression severity increased, shimmer\(^5\) and aspiration increased, and pitch variance and velocity tended to decrease (though not significantly). Horwitz et al. (2013) replicated the shimmer findings and additionally found that as depression increased, jitter\(^6\) increased and speech rate decreased. Trevino et al. (2011) took a less strictly acoustic route; they investigated phone rates as a potential informative feature for depression. Using an automatic phoneme recognizer, they obtained phone boundaries and durations in order to

---

\(^5\)The variability in frequency from one cycle to the next.

\(^6\)The variability in amplitude from one cycle to the next. Both jitter and shimmer are measures of vocal stability.
achieve a more localized rate. They determined that HAMD sub-measures of depression severity (discussed in Section \[2.2\]) correlated significantly with particular phonemes, and that as depression severity increased, speech rate (measured by phone rate) slowed.

Present research has very much moved in the direction of identifying robust and reliable features that can be used by algorithms to classify speech samples into depressed and non-depressed groups. Successful features from these studies include \(F_0\), energy, speaking rate, acoustic spectral features, and pitch. The following studies specifically involve the machine learning and automatic classification of depressed speech.

After compiling a corpus of depressed and control speech samples from various data sets, France et al. (2000) calculated \(F_0\), amplitude, formant, and power spectral density statistics. The found that classifiers that used multiple features were more successful than those that used single parameters. Moore et al. (2003) also extracted \(F_0\), energy, and speaking rate from a short story read by both depressed patients and controls, then used a Gaussian mixture model\(^7\) to classify the utterances into depressed and non-depressed classes. Results indicated that statistics related to \(F_0\) were the most informative, although these results contradict those found by Balsters et al. (2010), who did not find significant \(F_0\) differences between depressed and control subjects. Using a binary (depressed/not depressed) machine learning scenario, Alghowinem et al. (2013a) classified both read and spontaneous speech samples based on acoustic features like intensity, energy, voice quality, and pitch. They found that spontaneous speech was classified more accurately more often. Also, as they predicted, some features were more informative; jitter, MFCC, and energy features were found to be superior to pitch and voice quality features. When utilizing these features, an additional study by Alghowinem et al. (2013b) determined that in a binary scenario for spontaneous speech, a hybrid Gaussian mixture model and Support Vector Machine\(^7\) provided the best classification overall.

Many of the previously mention studies additionally investigated gender differences, furthering

\(^7\)A type of classifier in machine learning.
the research initiated by Ellgring & Scherer (1996). France et al. (2000) found that depressed female speech was generally characterized by spectral flattening. In depressed and suicidal men, they discovered lower F0 range and elevated formants. In the Moore et al. (2003) study, energy was found to be an important feature in the female group. Specifically investigating speech in depressed adolescents, Low et al. (2010) used Gaussian mixture models to classify speech based on Mel frequency cepstral features, the Teager energy operator, and pitch, formant, energy, jitter, and spectral features. They found that gender-based models outperformed gender-independent models, reiterating the observation of gender differences by Ellgring & Scherer (1996). For males, the addition of acoustic features to the Teager energy operator features increased accuracy, but this was not the case for females.

Despite all of this research, it is still unclear which features are best or most useful for differentiating depressed from normal speech. Many features have been proposed and investigated, the most prominent and common of which are acoustically-based, like pitch, energy, speech rate, pause time, and formant features. Trevino et al. (2011) began to move toward more theory-dependent features with phoneme identification. Continuing to bridge the gap between acoustic and linguistic analysis, this thesis explores a novel feature for depression detection: speech rhythm.

1.4 Speech Rhythm

Rhythm in speech, unlike most other types of rhythm, has little to do with periodicity and repetition; instead, it is the “systematic patterning of sounds in terms of timing, accent, and grouping” (Patel 2007). The notion of speech rhythm has historically been regarded as a differentiating factor between two groups of languages, so-called stress-timed and syllable-timed languages, resulting in rhythm typologies. Early theories that sought to explain these differences via isochrony were later disproven. Presently, rhythmic differences are largely attributed to phonotactic differences across languages and are measured by a set of consonant- and vowel-based metrics. The history of speech rhythm, its current theories and literature, and its criticisms will be discussed in this section.
1.4.1 Early Theory

The first rhythmic distinction between groups of languages is credited to Lloyd James (1940). In his training manual for British telephone operators in WWII, he advised that well-spoken English is similar to well-signdale Morse code: accented syllables must be clearly emphasized and evenly spaced because contrastive signals are more easily interpreted by the receiver. He stated that the “regular occurrence of accented syllables in our speech produces its characteristic rhythm” (25), and proposed that speech rhythm occurs in two variations. He introduced the first type as “Morse-code” rhythm languages, citing English, Arabic, and Persian, and referencing their more variable and less predictable unit durations. The second type he deemed “machine-gun” rhythm languages, such as French and Telugu, referencing their evenly timed, rapidly produced units. Thus was created the Morse-code and machine gun paradigm.

A few years later, Pike (1945) formalized the two groups as stress-timed and syllable-timed rhythms. At this time, the idea of isochrony, or the division of time into equal units by languages, was prevalent. Assuming that languages were isochronous, Pike aimed to determine which units were the basis for the isochrony in each language group. He defined a single rhythm unit as “a sentence or part of a sentence with a single rush of syllables uninterrupted by a pause” (34), and explained that in English, the length of time between the prominent, or stressed, syllables of successive rhythm units is roughly equal. He proposed that in order to maintain this consistent timing in longer rhythm units, syllables are “crushed together” and uttered more quickly in order to fit into the given time slot. Because the timing of English is therefore based on the rough periodicity of recurrent prominent/stressed syllables, Pike named it a “stress-timed” language. He contrasted this with languages like Spanish, whose syllable units (rather than stressed syllables) occur at generally uniform intervals, which he named “syllable-timed” languages. For a visual representation of this difference, see Figure 1.1. Because there is no pressure to constrain multiple syllables into a single timing unit, Pike
determined that syllable-timed languages have less need, and are less likely to, shorten and modify syllables or vowels. This was one of the first observations of specific differences between the two groups.

Coming from a different perspective, Abercrombie (1967) provided a physiological description for the existence of the two rhythm types. While his theory is interesting, it was disproven in the same year it was published and is given very little practical merit. Using Stetson (1951)’s speech production explanation as a foundation, Abercrombie suggested that an exhalation consists bursts of air, and that an individual chest burst, or pulse, is responsible for the formation of one syllable. Under this theory, stressed syllables result from reinforced chest pulses, which are produced by more forceful muscle contractions. He hypothesized the isochronic differences put forth by Pike (1945) were physiologically based, such that in syllable-timed languages, chest pulses recurred at equal intervals, while in stress-timed languages, rhythm is instead based on the reinforced chest pulses. Both types of languages achieve isochrony, but by different means.
1.4.2 Disproving Isochrony

Upon further examination, many of the assumptions underlying early speech rhythm theory were refuted. Abercrombie’s chest-pulse explanation was invalidated when, through a series of electromyographic experiments, Ladefoged (1967) determined that in conversational English, abdominal muscular activity only occurred in the final portion of a long utterance. It was therefore impossible for reinforced chest pulses to occur with every stressed syllable. Ohala et al. (1979) additionally provided evidence against the idea of chest-pulse-based isochrony, observing that active changes in lung volume only consistently occurred with heavily stressed syllables. Consequently, the chest-pulse theory of speech, and rhythm, was abandoned.

Although chest-pulse-based isochrony had been disproven, the idea of isochrony still existed as a potential underpinning of speech rhythm. Lehiste (1977) decided that factors influencing both production and perception of speech needed to be considered to determine whether English is isochronous. After analyzing recordings of read sentences, she determined that the size of inter-stress intervals varied quite a bit, demonstrating that English was, in fact, not isochronous. However, she posited that some of the differences in duration may be imperceptibly small, and therefore from a perceptual perspective, the intervals would be isochronous. Lehiste additionally examined differences in human judgments on speech and non-speech stimuli. She determined that subjects had more difficulty identifying duration variabilities in speech than in non-speech, indicating that isochrony may be a language-bound perceptual phenomenon. Her work suggested that, because humans tend to impose rhythm on sequences, isochrony could be largely perceptual, which calls into question the attempted differentiation of languages into rhythm classes on the basis of isochronic disparities.

\(^8\)Ohala et al. (1979) do not provide a definition for this term, but it likely refers to stressed syllables that are given additional emphasis.
1.4.3 Phonotactic Differences

In a pivotal paper, Dauer (1983) provided additional evidence against isochronic differences, showing that the inter-stress intervals maintained roughly the same average duration in English, Thai, Spanish, Italian, and Greek. Because these languages belong to different rhythm classes, the fact that their inter-stress intervals converged on an average duration suggests that perceived rhythmic differences were not produced by isochronous syllable units. This, plus the evidence against the previously held isochrony theory described above, led her to introduce a new explanation for the existence of the distinction between rhythm classes: phonotactics.

Dauer put forth that perceptual rhythmic differences in languages are due to differences in the languages themselves—specifically differences in syllable structure and the existence of vowel reduction processes. She explained that stress-timed languages typically permit a greater variety of syllable structures, which leads to more variation in syllable length and contributes to the perception of irregularity. Syllable-timed languages typically restrict onsets and codas to one or two consonants, while stress-timed languages often allow consonant clusters of three or more. Thus, there is more variation in potential consonant cluster length. She also points out that many stress-timed languages allow vowel reduction,⁹ which creates variations in vowel durations, increasing the amount of durational variability. Simply put, differences in the phonotactic structures of languages produce perceptions that they belong to one class or the other. Therefore, she concludes that a more accurate view of rhythm classes is that the languages lie on a spectrum (see Figure 1.2), rather than existing in two distinct groups; the phonotactics of a language, which determine the amount of durational variability of consonant clusters and vowels, establish its location on the stress-based continuum.

Dauer (1987) continued to develop this theory, moving toward an objective method of ascertaining a language’s stress-based-ness. She suggested a feature set, similar to distinctive feature theory,

---

⁹Defined by Dauer as “the centralization of unstressed vowels”. In some languages, this co-occurs with shortening processes. The fact that reduced vowels tend to be shorter is the relevant feature in this analysis.
where languages would be classified [+,-, 0] for features like syllable duration, syllable structure, pitch contours, vowel reduction, and consonant articulation. All of the feature values, with their definitions, can be found in Table 1.2. The greater the number of [+ values a language has for these features, the more stress-timed it is. Although this system was surely a step in the desirable direction of objectivity, many of the features are difficult to determine concretely. Their use requires the application of linguistic theory to determine the location of stress, for example, and are thus not entirely objective. The feature set has been occasionally utilized, as in Dimitrova (1997), but overall, the features never gained much popularity.

1.4.4 Rhythm Metrics

With the goal of creating entirely objective measures of rhythm, Ramus et al. (1999) used Dauer’s phonotactic explanation to develop the first of the modern rhythm metrics, segmenting sentences in various languages into vowel and consonant intervals in an attempt to capture the differences in consonant cluster length and vowel durations.

The process of segmenting speech into vowel and consonant intervals is represented in Figure 1.3. All adjacent consonants are aggregated into one interval, and likewise all adjacent vowels, without regard to syllable or word boundaries. In figure Figure 1.3, we see the phrase “great ideas” ([gɹeɪ-tɹdɪəz]) produced in isolation, and then segmented into phones. The initial complex onset [gɹ] becomes one consonant interval, [eɪ] becomes a vowel interval, and the [t] becomes a consonant interval. This continues until the entire utterance becomes a series of alternating consonant and vowel intervals.
### Table 1.2: A list of Dauer’s rhythm features for the classification of languages (Dauer 1987)

<table>
<thead>
<tr>
<th>Category</th>
<th>Feature</th>
<th>Value</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Length</strong></td>
<td>Duration</td>
<td>+</td>
<td>Accented syllables, and especially accented vowels, are regularly longer than unaccented syllables (by 1.5 or more) (e.g. English, Serbo-Croatian)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-</td>
<td>Accent does not affect the length of syllables, or the language has no accent (e.g. Japanese, Yoruba)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>Accented syllables are slightly longer than unaccented syllables (e.g. Spanish, Greek)</td>
</tr>
<tr>
<td><strong>Syllable Structure</strong></td>
<td></td>
<td>+</td>
<td>The language has a variety of syllable types (both heavy and light syllables with many different possible syllable structures), and heavy syllables tend to be accented, whereas light syllables tend to be unaccented (English, Arabic)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-</td>
<td>There are a very limited number of syllable types (predominantly CV or CVC), and accent and syllable weight are independent. There may be active processes such as final cluster simplification, epenthesis, or liaison to break up or prevent the formation of unusually heavy syllables (Spanish, French)</td>
</tr>
<tr>
<td><strong>Quantity</strong></td>
<td>+</td>
<td>Quantity distinctions, if present in the language, are only permitted in accented syllables; in unaccented syllables they are neutralized (only short) (some Arabic dialects)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>Quantity distinctions are permitted in both accented and unaccented syllables. Restrictions on quantity are not conditioned by accent (Hungarian, Finnish)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>All quantity distinctions occur in accented syllables, but only a small subset can occur in unaccented syllables (Estonian)</td>
<td></td>
</tr>
<tr>
<td><strong>Pitch</strong></td>
<td>Intonation</td>
<td>+</td>
<td>Accented syllables are turning points in the intonation contour. Pitch (usually high or changing) correlated with accent, but the actual pitch contour depends on the position in the utterance and the intonational meaning. Emphasis or contrast affects primarily the accented syllable (English, Greek)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-</td>
<td>Intonation and accent are independent; there may be a negative correlation of pitch and accent. Relative pitch patterns may be consistent with respect to the word regardless of its position in the utterance or intonational meaning. Emphasis may affect unaccented syllables or be achieved by other means (French, Japanese)</td>
</tr>
<tr>
<td><strong>Tone</strong></td>
<td>+</td>
<td>Tones, if present in the language, only exist on accented syllables; unaccented syllables are atonal (Swedish)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>Tones are present on all syllables or all syllables with a particular structure, regardless of accent. If there are sandhi rules, they are not related to accent (Yoruba)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>Tones are fully developed on accented syllables, but they are neutralized or subject to numerous changes (sandhi rules) in unaccented syllables (Thai)</td>
<td></td>
</tr>
<tr>
<td><strong>Quality</strong></td>
<td>Vowels</td>
<td>+</td>
<td>The maximal vowel system exists in accented syllables; vowels in unaccented syllables tend to be reduced or centralized (especially open vowels) (English, Swedish)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-</td>
<td>There is the same vowel system and similar articulation in all syllables. If elision or devoicing processes exist, they affect accented and unaccented vowels equally and are determined by phonetic environment rather than accent (Spanish, Japanese)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>The unaccented vowel system is smaller than that of accented vowels, but unaccented vowels are not necessarily centralized. There may be processes of devoicing or raising which occur only to unaccented vowels (Russian, Portuguese)</td>
</tr>
<tr>
<td></td>
<td>Consonants</td>
<td>+</td>
<td>Consonants are more precisely articulated in accented syllables, and some may have special reduced allophones (e.g. syllabic consonants, loss of aspiration) or be subject to neutralizations in unaccented syllables (English, Thai)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-</td>
<td>All consonants have the same articulation regardless of accent. Consonantal allophones are not conditioned by accent (French)</td>
</tr>
<tr>
<td><strong>Function of accent</strong></td>
<td></td>
<td>+</td>
<td>Accent can occur in different positions in a word (accent is &quot;free&quot; or free over a range) and is an integral part of the word shape for recognition. Moving the accent could result in a new word with a different meaning (English, Spanish, Swedish, Russian)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-</td>
<td>There is no word-level phonological accent; no one syllable consistently stands out over others in a word. Accent can be moved for stylistic or emotional reasons (in a language with a phrasal accent), but moving the accent does not result in a change in referential meaning or the establishment of new word boundaries (Yoruba, French)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>Accent can occur only in one position in a word (accent is &quot;fixed&quot;, typically on the first syllable). Moving the accent or adding an accent could result in the formation of a new word boundary (Hungarian)</td>
</tr>
</tbody>
</table>

After segmenting speech into consonant and vowel intervals, to capture the phonotactic differ-
REFERENCES identified by Dauer, Ramus et al. (1999) calculated the total vowel duration in an utterance (%V),\(^{10}\) the standard deviation of vowel interval lengths (\(ΔV\)), and the standard deviation of consonant interval lengths (\(ΔC\)). %C was not calculated because it would provide the same information as %V. To test the efficacy of these measures for distinguishing between rhythm classes, Ramus et al. plotted the measures for sentences in various languages, and found that %V plotted against \(ΔC\) yielded a plot that separated languages into groups that reflected rhythm classifications (see Figure 1.4). They therefore concluded that these two measures were able to capture rhythmic differences. Thus, %V and \(ΔC\) were touted as the first acoustically based, and therefore objective, rhythm metrics.

Grabe & Low (2002) believed that, while Ramus et al. (1999)’s metrics were headed down the right track, standard deviations of consonant and vowel interval lengths were not sufficient, and a more localized measure needed to be developed. Although the Ramus et al. (1999) metrics encom-

\(^{10}\)Metrics used in this analysis are underlined upon introduction.
passed a global measure of variability, they did not capture differences between adjacent intervals. Therefore, continuing in the vein of vowel and consonant intervals, Grabe & Low (2002) proposed calculating a Pairwise Variability Index, which aims to capture variability between successive intervals. A raw version exists (rPVI) which is used for consonants, as well as a normalized version (nPVI), which is used for vowel intervals, and controls for speech rate.\footnote{Equations for the PVI\textsc{es} can be found in Table 2.1.} For an example of this calculation, consider Figure 1.3. For the nPVI, the first and second and second and third V intervals would be compared, and likewise for the C intervals in rPVI. Grabe & Low (2002) acknowledged that the consonant interval PVI measure could be normalized, but noted that the value is affected by both speaking rate and syllable structure differences, which are difficult to separate. They there-

---

**Figure 1.4:** The plot of %V by Δ C from Ramus et al. (1999), demonstrating a separation of languages into their rhythm classes.
fore did not delve into details about consonant rate normalization and left consonant interval PVI measures unnormalized.

As did Ramus et al. (1999), Grabe & Low (2002) plotted their measures against each other, and found that rPVI and nPVI produced a similar, and they proposed, superior, separation of languages into their respective classes (e.g. Dutch, German, and English grouped together, while French and Spanish grouped together). Though some languages that are less clearly syllable- or stress-timed were classified differently by nPVI than %V, Grabe & Low (2002) asserted that the use of the PVIs is comparable to the use of $\Delta C$ and %V. They also noted that their results do not support the existence of a strict categorical distinction between stress- and syllable-timed languages, providing further support for Dauer (1983)’s stress-based continuum.

Barry et al. (2003) subsequently introduced a new PVI measure: PVI-CV. As motivation for the new metric, they argued that nPVI, whose normalization is intended to correct for speech rate, actually reduces stress- and accent-dependent and phonological length differences, which lie at the core of the syllable/stress-timed distinction. They also contended that PVIs fail to maximize their potential advantages over the Ramus et al. (1999) metrics because consonant and vowel intervals are analyzed separately, therefore ignoring the combined effects of consonant and vowel intervals on rhythmic perception. Barry et al. (2003) claimed that PVI-CV fully maximizes the PVI potential by calculating the variability of a combined consonant and vowel interval with the subsequent consonant-vowel interval pair in the utterance. For an example of how to segment a string into CV intervals, refer to Figure 1.5, which continues from the initial segmentation of “great ideas”. PVI-CV is calculated in the same manner as rPVI, with the first and second, second and third, and third and fourth intervals being compared.

The final canonical rhythm metric, Varco $\Delta C$, also commonly referred to as VarcoC, was proposed by Dellwo (2006). Based on previous research that $\Delta C$ varied in a manner dependent on speech rate (Dellwo & Wagner 2003), he put forward Varco $\Delta C$, aiming to measure variability while simultane-
1.4. SPEECH RHYTHM

Figure 1.5: The process of segmenting the phrase “great ideas” into CV intervals.

ously taking into account speech rate, which differs among languages. His metric employs a variation coefficient (Varco) and normalizes the variability by the mean consonant interval length for a given utterance. He confirmed its utility in a study with German, English, and French, ultimately determining that VarcoC separated languages into their rhythm classes more cleanly than ΔC.

Due to their ability to sort languages into their determined rhythm classes, the above-mentioned metrics are currently regarded as the most successful measurements of speech rhythm. Although a few other metrics have been proposed (e.g. YARD (Wagner & Dellwo 2004)), the consonant and vowel interval measures remain the most popular, likely because they do not depend on language-dependent measures like syllable, morpheme, or phrase boundaries.

1.4.5 APPLICATIONS OF THE RHYTHM METRICS

Since their introduction, researchers have used the rhythm metrics with varying degrees of success. They have applied the metrics to an assortment of languages to test their efficacy and cross-language
robustness.

For example, Gibbon & Gut (2001) found that nPVI was able to differentiate between British and Nigerian English and between British English and Ibibio, a Nigerian tone language in the Niger-Congo family. Using a slightly different, though fundamentally identical pairwise variability calculation, applied to both syllable and vowel durations, they compared utterances in British English, Ibibio, and Nigerian English. The calculations of PVI on syllables showed that Ibibio had more consistent syllable durations than the British English, supporting its syllable-timed classification. In a task that involved retelling a story, the PVI calculated on vowels was much higher for the British English speaker than for the Nigerian English speakers, indicating that Nigerian English vowel length varied less and suggesting that it is slightly more syllable-timed than British English. Given the resulting differences, they concluded that PVI is a valuable measure that is able to capture variations.

By applying the metrics to Urama, a Papuan language of New Guinea, Brown & Mandal (2013) confirmed that the metrics are able to capture the phonotactics of non-Western European languages. They obtained sentences in Urama, which has strict phonotactics; it only allows open syllables, prohibits consonant clusters, and does not reduce vowels. Urama had low values for consonant-based metrics as compared to English, Spanish, and Italian, reflecting its high proportion of vowels, and they concluded that generally, its scores followed the pattern of syllable-timed languages, as was expected from its phonotactics.

As a way of measuring variation from normal speech in a patient with Foreign Accent Syndrome (FAS), Yoon & Humphreys (2012) also employed the rhythm metrics, calculating %V, ∆C, ∆V, and the PVIs. Often in FAS, English speakers tend toward a more syllable-timed rhythm with no vowel reduction and unusually equal syllable durations, which were observed in the patient. They found that compared to the Buckeye Corpus baseline, the patient had significantly lower %V and nPVI values and higher ∆C and rPVI values, accurately capturing the deviant speech characteristics described as more syllable-timed.
Also investigating accented English, Mori et al. (2014) compared speech from American English speakers and Japanese English speakers. They found that when reading identical sentences, American English speakers had significantly higher standard deviations for vowel durations, reflecting a higher rate of reduction and a generalized difference between the two. The results indicated that the metrics were able to capture a production difference between native and L2 English.

Using all of the metrics addressed in this analysis in addition to others, Loukina et al. (2011) explored the metrics’ combined discriminatory power and found that a combination of metrics proved informative. Using automatic segmentation, they used machine classification algorithms to classify utterances in five languages. They found that when separating all five languages at once, only three measures were needed to achieve maximum accuracy and that many combinations of three metrics performed similarly. The most successful metrics were ratio measures like %V, normalized vocalic measures like VarcoV (the vowel interval parallel to VarcoC) and nPVI, and normalized CV-based measures like YARD (Wagner & Dellwo 2004).

1.4.6 Criticisms of the Metrics

While the metrics have been used with some success, they, along with the notion of speech rhythm in general, have also been criticized. Specific concerns include the exclusivity of timing measurements, the effects of elicitation method, speaker differences, and lack of consistency.

A common criticism of the metrics is that although rhythm may arise due to an interaction of features like pitch and stress, as put forth by Dauer (1987), the metrics solely capture timing variabilities. They therefore do not afford a complete representation of rhythm.

Toward this point, a study by Barry et al. (2009) suggested that F0 is an influential feature in rhythmic perception. They determined that for German speakers, variations in F0 produced more perceptions of regularity than variations in duration. Additionally, for Bulgarian speakers, F0 was ranked equally highly as duration for factors influencing rhythmicity judgments. The near-equal
importance of \(F_0\) to interval duration suggests that an effective measure of rhythm would include additional features, not just timing.

Investigating the interactions of speech rate and rhythm metrics, Dellwo (2008) pointed out that speech rate may inform rhythm classification more than was realized. In his study, participants listened to non-speech stimuli derived from German and French sentences and were instructed to rate how regular the sequences sounded. Linear regression models showed that CV interval rate (a measurement of speech rate) predicted listener regularity ratings better than \%V, nPVI, or VarcoC. This therefore suggests that speech rate may be equally or more important than timing variabilities due to phonotactic differences, in which case a combined consideration of timing and rate would be most informative.

Researchers have also raised concerns about the fact that methodologies can affect metric performance. Arvaniti (2012) compared utterances in different languages and found that metric scores very highly depended on speech elicitation method. This was echoed by Wiget et al. (2010), who determined that the lexical content of sentences significantly affected the metric values. Stimuli and elicitation methods tend to vary from study to study, which may help to explain the lack of consensus about average metric values for various languages.

Furthermore, it has been shown that inter-speaker variabilities contribute to muddled or conflicting results. Wiget et al. (2010) found that like sentence content, individual speaker differences significantly affected metric results. In an analysis of multiple speakers of multiple languages, Arvaniti (2012) additionally found that speakers from the same language did not always group together, and many of the scores from different languages intermingled. She concluded that there existed large inter-speaker variability and furthermore determined that speaker inconsistencies are more random, making them difficult to predict.

In extreme cases, the metrics have failed to recognize durational differences that they were designed to capture. For example, while observing rhythmic differences between two dialects of En-
Rathcke & Smith (2015) found that while the metrics somewhat behaved in the expected manner, nPVI was unable to capture vowel length differences found between speakers.

Recently, Nolan & Jeon (2014) published a paper challenging the traditional view of speech rhythm as a whole, labeling it a metaphor. They reiterated that the metrics only identify contrastive timing differences, while languages in reality use additional properties like variation in intensity and pitch. They suggested that speech rhythm is a metaphor—something that we understand but that may not reflect reality—and asserted that because of this, the search for acoustic correlates of linguistic rhythm is a fruitless effort.

The metrics were created to measure speech rhythm by capturing differences in phonotactic structures and have been used successfully by various researchers. In general, though, a definition of speech rhythm requires nuance, and as Wiget et al. (2010) indicated, the metrics’ extreme sensitivities warrant a tempered interpretation of results. The metrics themselves likely do not provide the most complete or accurate representation of rhythm, and would be more appropriate as part of a compound feature that encompasses other important dimensions identified by Dauer (1987) (see Table 1.2). Although there is doubt that the metrics capture rhythm and can capably distinguish rhythm classes, they do measure timing and timing variation within speech, which is the relevant feature for this paper.

1.5 DEPRESSION, DYSARTHRIA & SPEECH RHYTHM

While speech rhythm has not yet been investigated in depression, its use has met some success in dysarthric speech. Dysarthria is a motor speech disorder that occurs when speech production organs and muscles are impaired by injury to the motor cortex of the brain. This injury can be caused by events like strokes or accidents, or result from neurodegradation in diseases like Huntington’s or multiple sclerosis (ASHA 2015). A commonly researched subtype of dysarthria, hypokinetic dysarthria, often occurs in Parkinson’s disease, and is a consequence of damage to the basal ganglia,
which are heavily involved in the regulation of movement and muscles. Altered speech production pathways in hypokinetic dysarthria produce affected speech, which has been characterized by features like monopitch, reduced stress, harsh voice quality, monoloudness, and short rushes of speech (Duffy 2013). These closely mirror many of the dimensions used by Darby et al. (1984) to characterize depressed speech (see Section 1.3 for more a more detailed discussion).

In fact, psychomotor retardation occurs in both depression and Parkinson’s, and their physical symptoms can resemble one another so much in the early stages that it can be difficult to differentiate between them (Flint et al. 1992). Some hypothesize that shared psychomotor retardation indicates a shared underlying neurological basis (Caligiuri & Ellwanger 2000), perhaps disturbances in the basal ganglia pathways (Sobin & Sackeim 1997). If Parkinsonian dysarthria and depression share some underlying basis, we may expect some of the rhythmic (timing) characteristics of hypokinetic dysarthric speech to be reflected in depressed speech, and they may share similar coordination abnormalities.

Liss et al. (2009) used speech rhythm metrics with machine classification in order to discriminate among normal and different types of dysarthric speech. They used discriminate function analyses to determine which metrics accounted for the most variance among groups and found that many vowel-based metrics (specifically VarcoV, ΔV, and %V), VarcoVC, and ΔC were most informative for distinguishing among controls and various types of dysarthrias. An additional analysis with the inclusion of speech rate revealed high correlations between speech rate and all of the previously mentioned metrics, except VarcoVC. This confirms previous findings that speech rate interacts with timing variabilities (Dellwo 2008). Liss et al. (2009)’s success with the metrics in dysarthria prompts investigation into use of the rhythm metrics in other disorders.

As demonstrated in this chapter, the physiological effects of depression motivate research into vocal biomarkers for depression. Some features have already been determined, and this analysis explores a potential new feature. Given similar speech effects in depression and hypokinetic dysarthria,
and the metrics’ success in classification of dysarthric speech, I will investigate the usefulness of 
speech rhythm metrics in depressed speech. This analysis does not seek to quantify rhythm explic-
itly in depressed speech, but rather explores whether these rhythm-inspired metrics capture salient 
speech timing trends. By constraining segmentation to consonant and vowel intervals, we should be 
able to observe any durational changes or shortening linked to vowel or consonant reduction (van 
Son & Pols 1999; Aichert & Ziegler 2004) that emerge in depressed speech.
This chapter begins with a description of the rhythm metrics used in this analysis. It then discusses the depressed speech data, including patient demographics, speech tasks, and depression severity scores. The subsequent section explains the automatic segmentation process, and the final section details the calculation of the metrics and statistical analysis of the metric values.

2.1 Speech Rhythm Metrics

In this work, the measurements calculated were drawn from rhythm metrics and other features that have been shown to be related to depression. The canonical rhythm metrics, as described in Section 1.4.4, were each used, as they were in Liss et al. (2009), who found them to be useful in the
differentiation of normal and dysarthric speech. Additionally, metrics to measure features found in studies on depression to be relevant in depressed speech were calculated. These additional features were calculated with the expectation that their results would mirror previous findings in depression. Total sample duration (total time) and silence time, which includes pauses, were measured based on results from Ellgring & Scherer (1996), Cannizzaro et al. (2004), and Mundt et al. (2007), which demonstrated that silence increased with higher depression severity. A measure of speech rate was also calculated, as speech rate was found to be informative by Ellgring & Scherer (1996), Cannizzaro et al. (2004), Mundt et al. (2007), Horwitz et al. (2013), and Trevino et al. (2011). In this analysis, speech rate was measured as CV interval rate (Dellwo 2008), which captures the frequency of CV intervals (refer to Section 1.4.4 or Figure 1.5 for an explanation of CV intervals). Although syllable rate is often used to approximate speech rate (e.g. Grabe & Low (2002)), CV interval rate offers the distinct advantage of maintaining independence from language-dependent syllable boundary determinations, and is thus used as a proxy for syllable rate. Given that total silence is explicitly measured by the silence metric, CV interval rate focuses exclusively on the rate of consonant and vowel intervals. For a full list of metrics used in this analysis, see Table 2.1.
<table>
<thead>
<tr>
<th>Metric</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total time</td>
<td>Total length of sample</td>
</tr>
<tr>
<td>Total silence</td>
<td>Total duration of silence in sample</td>
</tr>
<tr>
<td>% silence</td>
<td>Percentage of speech time comprised of silence</td>
</tr>
<tr>
<td>%V</td>
<td>Percentage of sample comprised of vowel intervals</td>
</tr>
<tr>
<td>(\Delta V)</td>
<td>Standard deviation of vowel interval durations</td>
</tr>
<tr>
<td>(\Delta C)</td>
<td>Standard deviation of consonant interval durations</td>
</tr>
<tr>
<td>VarcoV</td>
<td>Standard deviation of vowel intervals divided by the mean vowel interval length (x100)</td>
</tr>
<tr>
<td>VarcoC</td>
<td>Standard deviation of consonant intervals divided by the mean consonant interval length (x100)</td>
</tr>
<tr>
<td>nPVI</td>
<td>Normalized Pairwise Variability Index: (\sum_{k=1}^{m-1}</td>
</tr>
<tr>
<td>rPVI</td>
<td>Raw Pairwise Variability Index: (\sum_{k=1}^{m-1} d_k - d_{k+1}/(m - 1)) (Grabe &amp; Low 2002)</td>
</tr>
<tr>
<td>rPVI-CV</td>
<td>Raw Pairwise Variability Index of successive CV intervals</td>
</tr>
<tr>
<td>CV interval rate</td>
<td>number of CV intervals produced per second (excluding silence and pauses)</td>
</tr>
</tbody>
</table>

Table 2.1: A list of the rhythm and other metrics, with their definitions, calculated in this analysis.
2.2 Data

The data used in this analysis was originally collected by Mundt et al. (2007) (see Section 1.3 for a discussion of this study) for use in a longitudinal study. Throughout the study, they collected speech data from depressed patients in order to track and explore acoustic changes during depression. The total dataset includes weekly recordings from 35 patients (20 female and 15 male), who were clinically diagnosed with depression at the onset of the study. During the study, all participants followed individual treatment regimens consisting of psychotherapy, pharmacotherapy, or a combination of the two. Additional demographic data included the gender and age of each patient, which ranged from 20 to 68. The age of one female participant was unavailable, and she was thus excluded from any analysis involving age.

Speech was collected from each subject in two ways: (i) through a telephone call made from outside, often in the participant’s home and (ii) in a clinician’s office via the office telephone. Every week, each participant called a telephone number and, prompted by an automatic system, completed a series of tasks (all were recorded at an 8kHz sampling rate). The in-office recordings were made every two weeks, during a meeting with a clinician. At these visits, participants made their recordings over the office telephone, following the same procedure. The speech was recorded directly through the telephone, therefore bypassing additional telephone channels and resulting in clearer recordings of higher quality. The office recordings existed for 28 of the 35 patients on weeks 0, 2, 4, and 6, totalling 112 samples for each speech task. For this analysis, the higher quality office recordings were used. The at-home recordings often included a great deal of background noise, so they were excluded in order to allow for the most accurate segmentation possible (discussed in the following section) and to avoid telephone-channel effects. Furthermore, this particular subset of the data was used in previous analyses on the Mundt et al. (2007) data (e.g. Trevino et al. (2011), Quatieri & Malyska (2012), Horwitz et al. (2013)), and is used in this analysis to maintain consistency.
During each phone call, participants performed various speech tasks which included counting from one to twenty, reciting the alphabet, a diadochokinetic task (repeating /pə tə kə/), sustaining vowel sounds, and reading *The Grandfather Passage* (Riper 1963)\(^1\) (included in Appendix A), which is a common task in speech evaluation. As an additional task, subjects were instructed to reflect on their past week in terms of their mental, physical, and general functioning, responding to each of the prompts for thirty seconds.

For this analysis, the free speech, alphabet, and *Grandfather Passage* speech samples were used. The free speech task captured speech that was entirely spontaneous. This was an important elicitation method to include, as Alghowinem et al. (2013a) found that spontaneous speech was classified accurately more often than read speech. In contrast to the free speech, the alphabet task provided a sample of memorized, mostly CV, monosyllabic speech. The counting task additionally provided memorized recitation, and /pataka/ task contained CV syllables, but the alphabet task would allow for exploration of recitation effects like singing, which are beyond the scope of this thesis. The *Grandfather Passage* task was longer, required reading, and included a mix of consonant cluster structures, which is more representative of English than CV syllables. These three tasks provide a diverse collection of speech production possibilities, allowing for the consideration of the effects of elicitation method.

For the free speech task, three prompts existed, and the first 30 seconds of the first prompt, which involved reflection on mental functioning, were used. Although participants were instructed to speak about the given topic for 30 seconds, some exceeded that length and some spoke for a shorter time. The shorter samples were used in their entirety while the longer samples were truncated at the pause closest to the 30-second mark.

Also included in the Mundt data are severity scores from two different depression mood scales:

\(^1\)For a more detailed history of the authorship of *The Grandfather Passage*, refer to Reilly & Fisher (2012).
the Quick Inventory of Depressive Symptomatology (QIDS) (see Appendix C.2) and the Hamilton rating scale for depression (HAMD) (see Appendix C). Patients completed self-report QIDS and HAMDS via a touch-tone system during every phone call, and the clinician also completed a HAMD evaluation during the biweekly consultations, totalling three depression scale scores for each office recording. The QIDS was created to be a short survey that addressed all diagnostic criteria for a major depressive episode in the DSM-IV (Rush et al. 2003), and is therefore intimately connected with current diagnostic procedure. The HAMD was created much earlier (Hamilton 1960) and has long been considered the gold standard for depression characterization. Both have advantages, and there is no clear determination about which is more accurate. Although all three scales (QIDS, HAMD-clinician, HAMD-self) were highly correlated, the scores are not identical (see Table 2.2 for correlation values). Therefore, each scale was analyzed independently. Because it is unclear which scale provides the most accurate measurement of depression severity, the results of one scale will not be weighted more heavily than another. Instead, metrics that correlate significantly with multiple scales will be considered more consistent and informative in depressed speech.

In order to allow for accurate segmentation of the data, which will be discussed in the following section, performances of alphabet and Grandfather tasks containing dysfluencies were discarded. I listened to each recording and any samples with false starts, corrections, incorrect words, or additional speech were removed. After discarding these samples, 107 alphabet samples and 60 Grandfather Passage samples remained. The free speech samples were transcribed by hand and therefore no samples were discarded due to dysfluencies; rather, dysfluencies were transcribed as proper speech.
CHAPTER 2. METHODOLOGY

2.3. AUTOMATIC SPEECH SEGMENTATION

The analysis thus included all 112 samples of free speech.

Based on suggestions by Ellgring & Scherer (1996) to investigate gender differences, which were bolstered by results from France et al. (2000), Moore et al. (2003), and Low et al. (2010), subjects were additionally divided by gender. Research has also indicated that both speech and reading rates slow with age (Ramig 1983), indicating that age may also be a potential confound. Visual analysis of the age distribution did not prompt any specific clustering or division among participants, so patients were split into the younger and older group by the mean, which was approximately 41. The female, male, older, and younger subgroups were analyzed independently of one another.

2.3. AUTOMATIC SPEECH SEGMENTATION

The majority of the rhythm metrics require segmentation of the speech into consonant and vowel intervals. Often, this segmentation is performed by humans trained in the task, as in Ramus et al. (1999), Grabe & Low (2002), and Liss et al. (2009). Because the future goal of this research is the development of automatic tools for the detection and classification of depression, the segmentation process must allow for a high volume of samples to be segmented extremely quickly. This is simply not possible with hand segmentation, but it is achievable via automatic segmentation of speech. Automatic speech segmentation works by identifying phone boundaries based on the spectral characteristics of speech. It has been shown that the boundaries produced by automatic segmentation fall within the range of boundaries produced by human segmentation (Wiget et al. 2010), so automatic segmentation is regarded as an accurate and suitable alternative.

The MIT Lincoln Laboratory Hydra system, which runs HTK (Cambridge University Engineering Department 2015) with a Hidden Markov Model trained on 100 hours of English StoryCorps data (StoryCorps 2016), was used to segment the Mundt et al. (2007) speech samples. Specifically, Hydra used automatic text alignment to determine phone boundaries, or the boundaries of individual sound units. Once Hydra determined the phone boundaries in a given sample, the boundaries
were returned in the format of a .phn (phone) file, which contains a list of all start times, stop times, and durations for each phone and pause in a sample.

In automatic text alignment, an English transcription of the speech is fed to a pronunciation dictionary, which then outputs a list of phones corresponding to the expected pronunciation of each word in the transcript. The list of phones informs the system, which then determines the boundaries (start and stop times) of each phone. Forcing alignment against a transcript increases accuracy over the alternative, speech decoding, because the system does not have to hypothesize which words were spoken.

When speech decoding is used, the system simultaneously identifies the word and its associated phone boundaries. This creates the potential for word identification errors, which then also results in phone identification errors. When there is no transcript available, the decoding process must be used. Therefore, in order to achieve maximal segmentation accuracy, the free speech samples were hand-transcribed. This, however, is a time-consuming process, increasing the appeal of content-controlled tasks like reading *The Grandfather Passage*, for which a single transcript can be applied. In both cases, though, the pronunciations expected by the system represent idealized or standard pronunciations. They therefore do not account for non-standard or deviant pronunciations which may occur in disordered speech.

### 2.4 Calculation of Metrics and Statistics

In order to calculate all of the metrics listed in Table 2.1, I wrote a Python script (see Appendix B) that takes a .phn file as input. Using the durations and phone labels from the .phn file, the script creates lists of C, V, silence, and CV intervals and their corresponding durations. From these intervals, the metrics are computed. Each .phn file was run through the script, which outputted the sample’s rhythm metric values. These values, along with participant demographic data, were aggregated for each task.
Then, statistical analyses were completed in MATLAB. For each task, Spearman correlations, which non-parametrically measure dependence between two variables, were calculated across all subjects and with all three depression scales. Correlations were additionally calculated within the female, male, older, and younger subgroups in each task. Additionally, Wilcoxon Mann-Whitney tests were calculated between gender and the metric values, and a Spearman correlation was calculated between age and the metric values for each task. For a correlation to be considered statistically significant, it necessarily achieved a p value of .05 or less.

In studies with multiple comparisons, a Bonferroni adjustment is often applied to account for correlations that are significant due to chance, or false positives. This involves adjusting the critical value (p value) based on the number of comparisons, in essence making it more difficult for a correlation to reach significance. Although a Bonferroni adjustment could be applied in this situation, it would increase the percentage of type II errors (false negatives) (McDonald 2014). Given the exploratory nature of the analysis, it is preferable to not discard any potentially significant results by applying such an adjustment.

Therefore, for each of the free speech, alphabet, and Grandfather tasks, there existed correlation values between each metric in Table 2.1 and each depression scale. This set of correlation exists across all participants and for each of the female, male, older, and younger subgroups.
The rhythm, silence, and articulation rate metrics were calculated on the phone files produced by the Hydra text-alignment automatic segmentation system. The three speech tasks (alphabet, Grandfather Passage, and free speech) were analyzed independently. Tasks were additionally assessed for potential confounding factors, namely gender and age. The subject for whom age was not recorded was excluded from all age analyses. The two age groups, younger and older, were split by the mean age of the subject group. Within each task, Spearman’s correlation coefficient was calculated between each of the metrics and depression severity scores.

The results will be presented as follows: each subject group, namely all subjects, males, females, older, and younger, will be discussed individually. Within each subgroup, the free speech, alphabet,
CHAPTER 3. RESULTS

and Grandfather tasks will be examined, in that order. Under each task, the relevant depression scale
scores will be noted. A summary of all results can be found in Table 3.8.

The Wilcoxon Mann-Whitney test results, which were calculated between gender and the metric
tables for each task, can be found in Table 3.1. The Spearman correlation results, which were calculat-
ed between age and the metric values for each task, can be found in Table 3.2.

<table>
<thead>
<tr>
<th></th>
<th>Free Speech</th>
<th>Alphabet</th>
<th>Grandfather</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total time</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Total silence</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>% silence</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>%V</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ΔV</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ΔC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VarcoV</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>VarcoC</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>nPVI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rPVI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PVI-CV</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>CV interval rate</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>

Table 3.1: All results from the Wilcoxon Mann-Whitney tests. Cells containing an ‘x’ indicate a significant difference ($p<.05$) in the means of the metric values between males and females.
### CHAPTER 3. RESULTS

#### 3.1. GROUP ANALYSIS: ALL SUBJECTS

<table>
<thead>
<tr>
<th>Metric</th>
<th>Free Speech</th>
<th>Alphabet</th>
<th>Grandfather</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total time</td>
<td>r = .2597</td>
<td>r = .4397</td>
<td></td>
</tr>
<tr>
<td>Total silence</td>
<td>r = .2194</td>
<td>r = .4087</td>
<td></td>
</tr>
<tr>
<td>% silence</td>
<td></td>
<td>r = .3647</td>
<td></td>
</tr>
<tr>
<td>%V</td>
<td></td>
<td>r = .2168</td>
<td></td>
</tr>
<tr>
<td>ΔV</td>
<td>r = .2168</td>
<td>r = .3002</td>
<td></td>
</tr>
<tr>
<td>ΔC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VarcoV</td>
<td>r = .2195</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VarcoC</td>
<td></td>
<td></td>
<td>r = -.2786</td>
</tr>
<tr>
<td>nPVI</td>
<td></td>
<td></td>
<td>r = .2850</td>
</tr>
<tr>
<td>rPVI</td>
<td></td>
<td></td>
<td>r = .2850</td>
</tr>
<tr>
<td>PVI-CV</td>
<td></td>
<td></td>
<td>r = .2814</td>
</tr>
<tr>
<td>CV interval</td>
<td></td>
<td></td>
<td>r = -.3300</td>
</tr>
</tbody>
</table>

Table 3.2: All significant (p < .05) correlations between age and the metric values.

#### 3.1.1. Free Speech

VarcoV was the only significant correlation in the free speech task (p < .05, r = .1964), and was correlated with the HAMD(self) depression scale. VarcoV was shown to increase as depression severity
increased, and these results indicate that vowel interval lengths became less variable. Although the correlation was significant, the magnitude was low, indicating a weaker correlation.

**Depression Scales**

- QIDS scores had no significant correlations.
- HAMD(self) scores correlated with VarcoV values.
- HAMD(clinician) scores had no significant correlations.

### 3.1.2 Alphabet

Figure 3.1 contains all significant correlations (p < .05) between metrics and depression scores found in the alphabet samples across all subjects. As depression severity increases, total silence time and % silence also increase, echoing previous results in demonstrating that more depressed patients have either longer or more frequent silences/pauses (or both). As depression severity increased, total time also increased, likely because of the increased silence time. Higher depression scores were inversely correlated with $\Delta C$, suggesting that more depressed patients had less variable consonant interval lengths. Only the silence metrics were significantly correlated with all three depression scale scores.

CV interval rate was not found to be significant, which indicates that the increase in total time was due to the increased silence and not slowed speech. Although the correlations were significant, the magnitudes were low, indicating weaker correlations.

**Depression Scales**

- QIDS scores were found to be significantly correlated with the total time and % silence metrics.
- The HAMD(self) scores significantly correlated with total time, total silence, and % silence.
- HAMD(clinician) scores significantly correlated with total silence and $\Delta C$. 
3.1.3 Grandfather Passage

Figure 3.2 contains all significant correlations ($p < .05$) between metrics and depression scores found in the Grandfather Passage samples across all subjects. Total time was also found to increase with increased depression severity, but it was correlated with a different depression scale than in the alphabet task. Silence was again found to increase with more severe depression, but only % silence was significant, and only with one depression scale. %V was found to increase with increased depression, suggesting that either vowels were uttered more slowly, consonants were shortened, or both. $\Delta C$ was again found to be inversely correlated with depression severity, but was also correlated with a different depression scale than in the alphabet task. VarcoC was found to decrease with depression severity for two of the scales, additionally indicating that consonant interval lengths became less variable. Again, CV interval rate was not found to be significant in this task. Although the correlations were significant, the magnitudes were low, indicating weaker correlations.

**Depression Scales**

QIDS scores correlated significantly with %V, $\Delta C$, and VarcoC values.

HAMD(self) scores had no significant correlations.

HAMD(clinician) significantly correlated with total silence, % silence, and VarcoC.
CHAPTER 3. RESULTS

3.2. GROUP ANALYSIS: FEMALES

**Figure 3.1:** Significant correlations in the alphabet task.

**Figure 3.2:** Significant correlations in the Grandfather Passage task.

3.2 **Group Analysis: Females**

<table>
<thead>
<tr>
<th></th>
<th>Free Speech</th>
<th>Alphabet</th>
<th>Grandfather</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Samples</td>
<td>64</td>
<td>62</td>
<td>30</td>
</tr>
</tbody>
</table>

*Table 3.4: The number of samples in each task across in the female subgroup.*
CHAPTER 3. RESULTS

3.2. GROUP ANALYSIS: FEMALES

3.2.1 Free Speech

In the female subgroup, a single significant correlation was found between total time and the self-report HAMD ($p<.05, r = .2738$). This indicated that total sample length increased with increased depression severity. Although the correlation was significant, the magnitude was low, indicating a weaker correlation.

**Depression Scales**

QIDS scores had no significant correlations.

HAMD(self) scores correlated with total time values.

HAMD(clinician) scores had no significant correlations.

3.2.2 Alphabet

Figure 3.3 contains all significant correlations found in the female subgroup. Total time and % silence increased and $\Delta C$ decreased with increased depression. The silence results mirror previous findings (e.g. Ellgring & Scherer (1996)), and the decreased $\Delta C$ indicates less variable consonant interval lengths with higher depression severity. Although the correlations were significant, the magnitudes were low, indicating weaker correlations.

**Depression Scales**

QIDS scores had no significant correlations.

HAMD(self) scores significantly correlated with total silence and % silence values.

HAMD(clinician) correlated with $\Delta C$. 

45
3.2.3 Grandfather Passage

Figure 3.4 contains all significant correlations (p<.05) between metrics and depression scores found in the Grandfather Passage samples across all female subjects. As depression increased, $\Delta V$ and $nPVI$ increased, indicating more variable vowel interval lengths with increased depression. $rPVI$ also increased, suggesting that consonant interval lengths also became more variable with increased depression. CV interval rate decreased with increased depression, demonstrating that when participants were more depressed, they spoke more slowly. The decreased speech rate is consistent with findings in previous literature. Although the correlations were significant, the magnitudes were low, indicating weaker correlations.

Depression Scales

QIDS scores correlated with $nPVI$ and CV interval rate values.

HAMD(self) scores had no significant correlations.

HAMD(clinician) scores correlated with $\Delta V$, $nPVI$, $rPVI$, and CV interval rate.

Figure 3.3: Significant correlations in the alphabet task in the female subgroup.
CHAPTER 3. RESULTS

3.3. GROUP ANALYSIS: MALES

![Graph showing significant correlations in the Grandfather Passage task in the female subgroup.]

**Figure 3.4:** Significant correlations in the Grandfather Passage task in the female subgroup.

### 3.3 Group Analysis: Males

<table>
<thead>
<tr>
<th></th>
<th>Free Speech</th>
<th>Alphabet</th>
<th>Grandfather</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of samples</td>
<td>48</td>
<td>45</td>
<td>30</td>
</tr>
</tbody>
</table>

**Table 3.5:** The number of samples in each task in the male subgroup.

#### 3.3.1 Free Speech

No significant correlations were found in the free speech task for the male subgroup.

#### Depression Scales

QIDS scores had no significant correlations.

HAMD(self) scores had no significant correlations.

HAMD(clinician) scores had no significant correlations.
CHAPTER 3. RESULTS

3.3. GROUP ANALYSIS: MALES

3.3.2 Alphabet

Figure 3.5 contains all significant correlations found in the male subgroup. Total time was found to increase with increased depression, and total silence and %silence were positively correlated with all three depression scale scores. %V and VarcoV were found to positively correlate with the QIDS and HAMD clinician scores. This again suggests increased silence with higher depression severity. The results also indicate either longer vowel or shorter consonant intervals and more variable vowel interval lengths with increased depression. Magnitudes of the correlations ranged from low to moderate, indicating weak to moderate correlations.

Depression Scales

QIDS scores correlated with total time, total silence, % silence, %V, and Varco V values.

HAMD(self) correlated with total silence and % silence.

HAMD(clinician) correlated with total silence, %silence, %V, and VarcoV values.

3.3.3 Grandfather Passage

Figure 3.6 contains all significant correlations (p<.05) between metrics and depression scores found in the Grandfather Passage samples across all male subjects. These results show increased silence with increased depression via the total silence and % silence metrics. % silence was significant across all three depression scales in this subgroup. The increased silence is consistent with previous findings both in the literature and in this analysis. As found in both the alphabet and Grandfather Passage tasks across all participants, ΔC decreased with increased depression, indicating less variable consonant interval lengths with increased depression severity. This is in contrast to the results from the female subgroup. Magnitudes of the correlations ranged from low to moderate, indicating weak to moderate correlations.
CHAPTER 3. RESULTS

3.3. GROUP ANALYSIS: MALES

Depression Scales

QIDS scores correlated with % silence and $\Delta C$ values.

HAMD(self) correlated with % silence and $\Delta C$ values.

HAMD(clinician) correlated with total silence and % silence values.

**Figure 3.5:** Significant correlations in the alphabet task in the male subgroup.

**Figure 3.6:** Significant correlations in the Grandfather Passage task in the male subgroup.
CHAPTER 3. RESULTS

3.4 GROUP ANALYSIS: OLDER

3.4.1 Free Speech

No significant correlations were found in the free speech task for the older subgroup.

Depression Scales

QIDS scores had no significant correlations.

HAMD(self) scores had no significant correlations.

HAMD(clinician) scores had no significant correlations.

3.4.2 Alphabet

A single correlation was found between CV interval rate and QIDS \((p<.05, r = .3023)\) in the alphabet task.

Depression Scales

QIDS scores correlated with CV interval rate values.

HAMD(self) scores had no significant correlations.

HAMD(clinician) scores had no significant correlations.

Table 3.6: The number of samples in each task in the older subgroup.

<table>
<thead>
<tr>
<th>Free Speech</th>
<th>Alphabet</th>
<th>Grandfather</th>
</tr>
</thead>
<tbody>
<tr>
<td>52</td>
<td>49</td>
<td>30</td>
</tr>
</tbody>
</table>
3.4.3 Grandfather Passage

Figure 3.7 contains all significant correlations ($p < .05$) between metrics and depression scores found in the Grandfather Passage samples across all older subjects. Total time, total silence, and % silence increased with increased depression severity, consistent with prior results. $\Delta V$ and nPVI increased with increased depression, and as in the female subgroup, this indicates more variable vowel interval lengths with more severe depression. Also echoing results found in the female subgroup, rPVI in increased with increased depression, indicating less variable consonant intervals with increased depression. CV interval rate decreased with increased depression across all three depression scales, indicating slower speech with more severe depression. This is consistent with both the previous literature (see Section 1.3) and the results from the female subgroup. Total time and CV interval rate were significant across all three depression scales. The effects of aging on speaking and reading rates (Ramig 1983) likely compounded the effects of depression in the older group, increasing total time and decreasing CV interval rate. Magnitudes of the correlations ranged from low to moderate, indicating weak to moderate correlations.

Depression Scales

QIDS scores correlated with total time, $\Delta V$, nPVI, rPVI, and CV interval rate values.

HAMD(self) correlated with total time, total silence, % silence, and nPVI values.

HAMD(clinician) correlated with total time, total silence, % silence, $\Delta V$, nPVI, rPVI, and CV interval rate.
CHAPTER 3. RESULTS

3.5. GROUP ANALYSIS: YOUNGER

Figure 3.7: Significant correlations in the Grandfather Passage task in the older subgroup.

3.5 Group Analysis: Younger

<table>
<thead>
<tr>
<th>Free Speech</th>
<th>Alphabet</th>
<th>Grandfather</th>
</tr>
</thead>
<tbody>
<tr>
<td>56</td>
<td>54</td>
<td>28</td>
</tr>
</tbody>
</table>

Table 3.7: The number of samples in each task in the younger subgroup.

3.5.1 Free Speech

No significant correlations were found in the free speech task for the younger subgroup.

Depression Scales

QIDS scores had no significant correlations.

HAMD(self) scores had no significant correlations.

HAMD(clinician) scores had no significant correlations.
3.5.2 Alphabet

Figure 3.8 contains all significant correlations found in the younger subgroup. % silence was found to positively correlate with both the QIDS and HAMD (clinician) depression scale scores. This echoes previous findings that silence increases with increased depression. Although the correlations were significant, the magnitudes were low, indicating weaker correlations.

**Depression Scales**

QIDS scores correlated with % silence values.

HAMD(self) scores had no significant correlations.

HAMD(clinician) correlated with % silence.

3.5.3 Grandfather Passage

Figure 3.9 contains all significant correlations ($p < .05$) between metrics and depression scores found in the Grandfather Passage samples across all younger subjects. %V was found to increase with increased depression, as in the analysis across all participants in the Grandfather task, suggesting an overall vowel interval lengthening or consonant interval shortening. $\Delta C$ was found to significantly decrease across all three depression tasks, indicating less variable consonant interval lengths, and replicating results found previously in this analysis. VarcoC also significantly decreased across all three depression scales, furthermore indicating less variable consonant interval lengths with increased depression. Magnitudes of the correlations ranged from low to moderate, indicating weak to moderate correlations.

**Depression Scales**

QIDS scores correlated with $\Delta C$ and VarcoC values.
HAMD(self) correlated with %V, ∆C, and VarcoC values.

HAMD(clinician) correlated with %V, ∆C, and VarcoC.

Figure 3.8: Significant correlations in the alphabet task in the younger subgroup.

Figure 3.9: Significant correlations in the Grandfather Passage task in the younger subgroup.
### Table 3.8

A summary of the results from this analysis. All significant correlations for each task and subgroup are listed, in addition to whichever depression scale scores the metrics significantly correlated with.

<table>
<thead>
<tr>
<th></th>
<th>All Subjects</th>
<th>Females</th>
<th>Males</th>
<th>Older</th>
<th>Younger</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Free Speech</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VarcoV: HAMD(s)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total time: HAMD(s)</td>
<td>Total silence: ALL</td>
<td>% silence: ALL</td>
<td>CV interval rate: QIDS</td>
<td>% silence: QIDS, HAMD(c)</td>
<td></td>
</tr>
<tr>
<td>%V: QIDS, HAMD(c)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VarcoC: QIDS, HAMD(c)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Alphabet</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total time: HAMD(s)</td>
<td>Total silence: ALL</td>
<td>% silence: HAMD(c)</td>
<td>CV interval rate: QIDS</td>
<td>% silence: QIDS, HAMD(c)</td>
<td></td>
</tr>
<tr>
<td>%V: QIDS, HAMD(c)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VarcoC: QIDS, HAMD(c)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Grandfather</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total silence: HAMD(c)</td>
<td>Total silence: HAMD(c)</td>
<td>CV interval rate: QIDS</td>
<td>% silence: ALL</td>
<td>%V: HAMD(s), HAMD(c)</td>
<td></td>
</tr>
<tr>
<td>%: QIDS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>nPVI: QIDS, HAMD(c)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rPVI: HAMD(c)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CV interval rate: QIDS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VarcoC: QIDS, HAMD(c)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

**Note:**
- VarcoV: HAMD(s) indicates the variance components of the HAMD(s) scale.
- CV interval rate: QIDS refers to the coefficient of variation for the QIDS scale.
- %V: HAMD(s), HAMD(c) denotes the percentage of variance explained by the HAMD(s) and HAMD(c) scales.
- All significant correlations are listed for each task and subgroup, in addition to the depression scale scores they are significantly correlated with.
In this chapter, the results from Chapter 3 will be interpreted. Each metric will be discussed individually, and general conclusions will be drawn, addressing the usefulness of the metrics in determining the existence and severity of depression based on a speech sample. From there, the chapter will move to a discussion of potential confounds and suggestions for future research, and will ultimately conclude with considerations of the current approach to speech rhythm and its implications for research.
4.1 Discussion of Correlation Results

Each metric from Table 2.1 was correlated with depression scale scores across multiple tasks and within various subgroups. The results, which are summarized in Table 3.8, will be discussed, addressing the performance of each metric individually. The non-rhythm metrics will be presented first, and then the rhythm metrics, in order of significance frequency, will be listed.

Figure 4.1 contains all of the metrics, displayed with the number of significant correlations each had with one of the three depression scales across all analyses. Since there is no basis for choosing one depression scale over the others, each scale is deemed a possible correlation. As a result, the highest possible number of significant correlations for each metric is 45, which would be achieved by a metric correlating significantly with every depression scale (3) in every task (3) in every analysis group (5). This provides a visual comparison of the metrics’ performances across the data.
The non-rhythm metrics will be presented first, and then the rhythm metrics, in order of significance frequency, will be listed.

**Total time.** Total time was correlated significantly with depression scale scores 6 times throughout the analysis, increasing with increased depression in each case. For total time to increase, either the amount of silence contained in a sample increased or speech rate decreased. Given the more frequent occurrence of increased silence in the data, the increase in total time likely occurred as an effect of increased silence. Specifically for the older group in the *Grandfather* task, speech rate also likely decreased, considering that reading rate slows with age (Ramig 1983).

**Total silence.** Total silence correlated with depression scale scores in 11 instances and was the second most frequent significant metric. This mirrors previous results finding that found that silence
frequency and duration increase with increased depression severity (Ellgring & Scherer 1996; Cannizzaro et al. 2004; Mundt et al. 2007).

% silence. % silence obtained the highest number of significant correlations of all the metrics, correlation with depression severity scores 15 times. This is unsurprising given the number of studies that have found silence to be indicative of depression severity (See Table 1.1). Although the increase in % silence could be indicative of shorter speech intervals, the significance of total silence suggests that silence itself increased.

CV interval rate. CV interval rate reached significance 6 times in the data, suggesting that it only mildly varied with depression severity. Previous studies have found speech rate to be affected by depression severity (Ellgring & Scherer 1996; Cannizzaro et al. 2004; Mundt et al. 2007; Horwitz et al. 2013; Trevino et al. 2011), and the same results were expected in this analysis. However, CV interval rate, the measure of speech rate used in this analysis, failed to reach significance in most cases. This seems to contradict previous findings that speech rate decreases with increased depression. A potential explanation of this is that the rate measurements in previous studies seem to have been calculated with the inclusion of silence, while this study excluded it. As silence was found to be a significant factor, it follows that speech rate measurements that include silence should also reach significance. The failure of CV interval rate to reach significance indicates that the rate of actual speech units was unaffected by depression severity. This prompts the question of how best to measure speech rate and careful consideration of what features are actually measured by different speech rate measurement methods.

ΔC. ΔC was the third most frequent significant metric, achieving significance in 8 instances in the data. Across the groups and tasks, ΔC decreased with increased depression severity. This suggests that in depressed speech, consonant intervals were less variable in length, and taken with the increase in %V, could have been shorter as well. The shorter and less variable consonant intervals may indicate that as depression increases, patients produce less precise and controlled articulatory
movements, which results in incomplete or co-articulated, and thus shorter, consonant cluster production. The decrease in motor precision could be the product of emotional effects (e.g. less effortful speech), psychomotor retardation, or an interaction of the two.

%V. %V was found to occasionally correlate with depression severity levels. It tended to increase with increased depression, which could indicate longer vowel intervals, shorter consonant intervals, or a combination of the two. Although Liss et al. (2009) found vowel-based metrics to be particularly informative in the differentiation of dysarthric speech, %V’s low frequency of significance does not reflect their findings and may indicate a difference between dysarthric and depressed speech.

VarcoC. VarcoC, although it occasionally reached significance, also did not prove to be the most informative feature for depressed speech. Like $\Delta C$, it decreased with increased depression severity, indicating less variable consonant intervals in more depressed subjects, but the rate normalization seems to have made it less predictive of depression severity than its non-normalized counterpart.

nPVI. nPVI seldom achieved significance in the data, correlating with depression severity level 5 times across the entire analysis. It tended to increase with increased depression, indicating more variable vowel interval lengths. However, its infrequent significance suggests that it is not very informative in depressed speech.

$\Delta V$. $\Delta V$ obtained very few significant correlations. This suggests that it is not very informative for depression severity, and like the %V results, does not replicate the findings of Liss et al. (2009).

VarcoV. VarcoV, like $\Delta V$, very infrequently reached significance. This suggests that vowel-based metrics may not be as informative for depression severity level as they were for Liss et al. (2009) in dysarthric speech.

rPVI. rPVI achieved even fewer significant correlations than $\Delta C$ and VarcoC, reaching significance 3 times in the data. Like the other two aforementioned consonant-based metrics, it decreased with increased depression, indicating less variable consonant interval lengths. However, it performed more poorly than the other consonant-based metrics and therefore would likely not provide
CHAPTER 4. DISCUSSION

4.1. DISCUSSION OF CORRELATION RESULTS

much additional information.

**PVI-CV.** PVI-CV performed the worst of all the metrics, failing to reach significance even once. This suggests that this metric is unaffected by depression severity level and would not be useful in the classification of depressed speech.

**Liss et al. (2009)**’s study served as one of the models for this analysis, and the similarity of depressed and dysarthric speech (see Section 1.5) provided motivation for use of the rhythm metrics in depressed speech. However, although in **Liss et al. (2009)**’s study the metrics were able to differentiate between normal and dysarthric speech, their performance with depressed speech was less successful. Many of the metrics that **Liss et al. (2009)** found to be informative were the weaker metrics in this analysis. For example, they found vowel-based metrics to be particularly successful at discriminating dysarthric and non-dysarthric speech. In this analysis of depressed speech, the correlations with those metrics were sporadic and relatively weak. Therefore, this may indicate that although depression and dysarthria have similar effects on speech, and may have underlyingly similar neurological bases, the two disorders affect speech differently, such that relevant measures in one may not be as relevant in the other.

Overall, these results suggest that the rhythm metrics may have promise for capturing timing differences in depressed speech. None of the metrics performed outstandingly, though, as applied to the **Mundt et al. (2007)**, and % silence, the metric that was significant the most often, only reached significance in one third of analyses. Of the rhythm metrics, ∆C appears the most promising. Its significance suggests a change in the production of consonants and consonant clusters in depressed speech, an area that has yet to be explored.

The relatively infrequent significant correlations, in addition to the low magnitudes of the correlations, suggest that timing differences, if they exist in depressed speech, were mildly captured by the rhythm metrics in this analysis. Therefore, these metrics may be best utilized as one dimension in a complex feature, compounded with other features like pitch and intensity measurements. Various
factors like data collection technique and segmentation errors, which will be discussed further in the next section, may have confounded results. Therefore, further investigation into these confounds is recommended, with specific focus on the ways that depressed speech differs from normal speech, and what those differences implicate for speech analysis methods. Additional research into potential production phenomena that are intimated by the metrics, like consonant cluster production disruption, is also recommended.

4.2 Potential Confounds

There are a number of potential confounds that may have affected these results and that should be considered in future research on the automatic detection of depression severity from speech samples.

Firstly, it must be considered that fatigue is often comorbid with depression, insofar as fatigue is a diagnostic criterion of Major Depressive Disorder (American Psychiatric Association 2013). Many of the speech features that can indicate fatigue (discussed in Section 1.1), like speech rate and pause time (Vogel et al. 2010), are also relevant in depression. As of now, we cannot extract fatigue from depression, and therefore we cannot be certain which speech changes are due to fatigue and which are caused by depression. An aspect of future research in depression should focus on identifying distinctive features for each so that analyses can specifically address speech effects caused by depression without inadvertently capturing effects caused by other conditions.

Since automatic segmentation requires discrete phone boundaries, another confound to consider is the process of segmenting continuous speech into individual phones. In general, this process is problematic because in natural speech, phones do not exist in isolation. By requiring discrete phone boundaries, co-articulation and anticipation effects, which occur constantly in fluid speech, are lost. Therefore, segmented speech does not completely reflect the reality of natural speech, which does not contain these clear-cut phone boundaries. Despite this issue, segmentation remains extremely useful and is necessary for many automatic means of detecting depression severity from speech.
CHAPTER 4. DISCUSSION

4.3 DIRECTIONS FOR FURTHER RESEARCH

Some studies of speech rhythm treat specific phonological situations differently, creating segmentation inconsistencies across studies. In many of the studies that utilized hand segmentation, particular allowances, like the inclusion of a word-initial glide in a vowel interval, were made (e.g. Grabe & Low (2002)). These phone-specific situations occur throughout the data, and although the method used in this analysis would not support these special-case classifications, it is possible to create a methodology that would. These allowances could increase the consistency of the metrics.

Further addressing automatic segmentation, text alignment applies an idealized pronunciation to speech, which may differ greatly from the speech of the depressed subjects. Participants likely did not produce speech that mirrored standard or expected pronunciations and therefore the phonetic representation of the word used by the system may not fully match reality. If a participant produced a non-standard pronunciation, the system would still look for the phones expected in the prescriptive standard pronunciation, increasing error rates for phone boundary and label identifications. This could be avoided by producing careful phonetic transcriptions for each speech sample, but this would be extremely time-expensive. This issue merits further research, as detailed in the following section.

4.3 DIRECTIONS FOR FURTHER RESEARCH

Based on the results of this thesis, and the potential confounds listed above, there are several clear directions for future research on the timing of depressed speech. These involve automatic segmentation considerations, identification of the best speech samples for analysis, and the exploration of the extent to which speaker differences influence the results of an across-group analysis.

As the future goal of this research is the development of automatic diagnostic and monitoring tools that could be used outside of a clinician’s office, research should also be conducted using lower-quality recordings. The higher quality in-office data from the Mundt et al. (2007) study was used for this analysis, and a similar experiment could be conducted on the at-home data, which was collected
as part of the same study but not analyzed here. Determination of telephone channel and noisy environment effects would create the potential for an exploration of automatic segmentation accuracy on lower quality data. A gold standard of hand-segmented data would be an important addition, allowing for investigations into accuracy effects due to quality reduction.

An additional gold standard of hand-segmented depressed and non-depressed speech would be ideal. Alignment systems, including the one used in this analysis, are generally trained on normal speech, and it is unknown how variations in speech due to depression would affect the segmentation. Hand segmentation of depressed speech would additionally provide more detailed information about the characteristics of depressed speech (including the effects of depression on speech segments, including effects on consonant clusters) and would allow for considerations of segmentation methodologies specifically designed to account for alterations that occur in speech due to depression.

A primary consideration is that the data used in a future analysis be best suited to reveal correlations between depression severity and timing effects. Alghowinem et al. (2013a) found that spontaneous speech was classified into depressed and non-depressed categories more accurately more often than read speech, but the results of this analysis would suggest otherwise. In this study, the free speech task obtained the fewest correlations between metrics and depression severity scores, and three subgroups did not have any correlations whatsoever. In fact, the samples of read speech (derived from Grandfather Passage task) achieved the most significant correlations. Thus, it is clear that not all speech samples are equal, and consideration of the content of the tasks and their effects on the metric values is warranted.

Since it has no controlled lexical content, the free speech task may not provide the best baseline for comparison between subjects. The rhythm metrics were designed to measure phonotactic variabilities, and they are extremely sensitive to differences in syllable structures in the speech samples, and the differences in the content in the speech samples derived from the free speech task likely con-
found the results. Additionally, during the task, subjects struggled to think of new information to convey and inserted pauses frequently as they attempted to respond to the prompt for thirty seconds, producing samples containing long pauses and fillers. Pauses and fillers can be expected in response to any prompt where a subject is asked to speak for a specific amount of time, providing another reason why speech samples from a free speech task are not ideal for this type of analysis.

The alphabet task, in contrast, is content-controlled, but participants produced the alphabet differently; some recited it as a list, others sang it as a song. Recitation method likely influenced silence, CV interval rate, and vowel duration measurements. Furthermore, the alphabet is not representative of the syllabic structures in a natural English sentence; it is composed of primarily CV syllables.

The Grandfather Passage task contained content-controlled sentences that reflected the variations in syllable structure allowed by English phonotactics. Reading styles may have contributed to differences between subjects. As the task was a story, sentences may have been affected by the previous and subsequent sentences, particularly in more dramatic readings, which would not have occurred if each sentence were produced individually. Considering this, it seems that single sentences produced in isolation may provide the most suitable data for rhythm metric analyses, as suggested by Wiget et al. (2010).

Additionally, since the rhythm metric values may be confounded by inter-speaker differences (Arvaniti 2012; Wiget et al. 2010), longitudinal studies of individual speakers should be conducted. Rather than an across-subject analysis, like the one in this thesis, a longitudinal study following individual speakers over an extended period of time may have more success. The comparison of data from multiple speakers may help to explain the inconsistent results of the metrics in this study, which aggregated data from 28 individuals of different ages and genders, and with different treatment protocols. The results of a longitudinal study would inform a study looking across subjects for the identification of the best timing metrics, allowing for compensation for individual differences. For each individual subject, it would also be ideal to obtain speech samples from a non-depressed
state to use as a baseline, providing information on the speaker’s base speech rate and pronunciation.

Finally, future research should aim to identify which of the three depression scales used in this analysis provides the most reliable measurement of depression severity possible to allow for simple and accurate evaluation. In terms of quantifying depression severity, it is unclear what rating method is best. The depression scales (discussed in Section 2.2) are highly correlated with one another, but are not identical. When a was metric significantly correlated with a depression scale score, it did not always correlate with all three, exemplifying the inconsistency in the depression scales. Consideration of just one scale may not provide a complete picture, but the use of three depression scales requires the administration of three different surveys. Future research may reveal that a weighted-average of the depression scores would be ideal.

4.4 Speech Rhythm and Timing Metrics

There exists debate about whether the metrics are truly able to capture rhythm (see Section 1.4.6) or simply quantify timing differences. The metrics likely capture aspects of rhythm, but are unable to account for the entire perceptual phenomenon that resulted in the identification of stress- and syllable-timed languages. It seems reasonable that a complete definition of rhythm includes features beyond just simple duration differences, and instead that speech rhythm is comprised of more than just timing differences or is due to more than just the phonotactics of a language. As a result, this thesis makes no claim about changes in the rhythm of depressed speech, but instead investigates timing differences using metrics that were originally created to measure rhythm.

Therefore, the research conducted in this thesis is independent of the outcome of researchers’ determination of the complete set of factors that determine the rhythm of a language. This analysis focuses on the timing of depressed speech, which many believe is a major factor contributing to a language’s rhythm, and which is well suited for automatic measurement. If features are identified that are better able to holistically capture rhythm, the question of rhythmic changes due to depres-
sion remains open. In fact, if future features compound multiple features, like durations and pitch, these compound features should be applied to depressed speech, as they would address multiple dimensions that have shown to be affected by depression (see Section 1.3). The current rhythm metrics provide mild potential for use on their own in depressed speech, but speech rhythm, or the perception of rhythm in depressed speech remains unexplored.
In this thesis speech rhythm metrics were explored as a potential informative feature in the automatic evaluation of Major Depressive Disorder. Automatic text alignment was used to segment a collection of recordings made by depressed patients over a period of six weeks. From the phone boundaries produced by the automatic text alignment system, consonant, vowel, and silence intervals were determined, and rhythm metrics were calculated. The rhythm metric values were correlated with various self-report and clinician-assigned depression severity score data available from each recording day to determine whether the values of any metrics co-varied with depression severity. In order to explore the effect of gender and age, which were potential confounds, each speech task was additionally divided into subgroups, which were analyzed independently.
CHAPTER 5. CONCLUSIONS

Silence was found to be correlated with depression severity in two content-controlled tasks. Additionally, consonant interval durations were found to become less variable in length with increased depression. This decreased variability points to a potential decrease in motor precision resulting in co-articulated or otherwise reduced consonant clusters (van Son & Pols 1999) with increased depression, a subject meriting further exploration in future studies. Although no metrics were entirely consistent, task differences, subjective depression scales, a lack of a normal control group, and inter-speaker variations contribute to confounded metric values.

These results prompt further research into the production of complex consonant clusters in depressed speech, as was explored in other contexts in Aichert & Ziegler (2004). Future analyses could also look at the variation in metrics for a single speaker over a longer period of time, as suggested by Wiget et al. (2010), or include additional features like pitch or intensity variations, which have been found to be relevant in depression and the perception of rhythm. The use of automatic segmentation also would allow for exploration of a silence feature, which could prove informative in depression evaluation. More research into the effects of noisy environments and non-normal speech on segmentation accuracy is recommended.

This analysis serves as a contribution to the establishment of features for use in the automatic detection of depression severity. The confounds encountered are relevant to speech analysis in other disorders, as is future research into compound feature development and segmentation methods. Several options are available for further exploration of the development of robust features for use in scalable automatic tools, which could be invaluable for providing effective, accessible aid to the ever-growing affected population.
The Grandfather Passage (Riper 1963):

You wished to know all about my grandfather. Well, he is nearly ninety-three years old. He dresses himself in an ancient black frock coat, usually minus several buttons; yet he still thinks as swiftly as ever. A long, flowing beard clings to his chin, giving those who observe him a pronounced feeling of the utmost respect. When he speaks his voice is just a bit cracked and quivers a trifle. Twice each day he plays skillfully and with zest upon our small organ. Except in the winter when the ooze or snow or ice prevents, he slowly takes a short walk in the open air each day. We have often urged him to walk more and smoke less, but he always answers, “Banana Oil!” Grandfather likes to be modern in his language.
The following Python script, written as part of this research, takes a .phn file as input, then creates consonant, vowel, and silence intervals. The intervals are used to calculate all rhythm metrics used in this analysis.

```python
#rhythmmetrics.py
import numpy as np
import pandas as pd
from pandas import Series

#read in phone file and create dataframe
fname = input('phn fname: ')
df = pd.read_csv(fname, sep=' ', header=None, names=['Start', 'Stop', 'Phone'])
df2 = df.copy()
df2.insert(2, 'Duration', df2.Stop-df2.Start)

#arpabet vowels

#insert new Category column, all values are a default 'C'
df2.insert(4, 'Category', 'C')

#change Category values based on Phone values
for i in range(len(df2)-1):
    if (df2.loc[i, 'Phone'] == 'sil') or (df2.loc[i, 'Phone'] == 'sp'):
```

71
    df2.loc[i, 'Category'] = 'SIL'
    elif df2.loc[i, 'Phone'] in ar_vowels:
        df2.loc[i, 'Category'] = 'V'

    # new column will hold cumulative intervals
    df2.insert(5, 'Interval', df2.Duration)

    # Compute intervals durations (collapse adjacent phones of same category into single intervals)
    for i in range(len(df2)-1):
        if df2.loc[i, 'Category'] == df2.loc[i+1, 'Category']:
            df2.loc[i+1, 'Interval'] = df2.loc[i, 'Interval'] + df2.loc[i+1, 'Interval']
        df2.loc[i, 'Interval'] = '----'

    # lists of interval duration values
    vint = []
    cint = []
    silint = []

    # get list of V intervals and insert into a column
    for i in range(len(df2)):
        if (df2.loc[i, 'Category'] == 'V') and (df2.loc[i, 'Interval'] != '----'):
            vint.append(df2.loc[i, 'Interval'])
    df2.loc[:, 'Vowels'] = Series(vint)

    # get list of C intervals and insert into a column
    for i in range(len(df2)):
        if (df2.loc[i, 'Category'] == 'C') and (df2.loc[i, 'Interval'] != '----'):
            cint.append(df2.loc[i, 'Interval'])
    df2.loc[:, 'Cons'] = Series(cint)

    # get list of SIL intervals and insert into a column
    for i in range(len(df2)):
        if (df2.loc[i, 'Category'] == 'SIL') and (df2.loc[i, 'Interval'] != '----'):
            silint.append(df2.loc[i, 'Interval'])
    df2.loc[:, 'Sil'] = Series(silint)
#get list of V and C intervals in their chronological order
df2['cvCat'] = ''
df2['cvDur'] = ''
cvcat = []
cvdur = []
for i in range(len(df2)):
    if (df2.loc[i, 'Category'] != 'SIL') and (df2.loc[i, 'Interval'] != '----'):
        cvcat.append(df2.loc[i, 'Category'])
        cvdur.append(df2.loc[i, 'Interval'])
    df2.loc[:, 'cvCat'] = Series(cvcat)
    df2.loc[:, 'cvDur'] = Series(cvdur)

#loop through list of durations and collapse CV intervals into a single duration, then, insert them into a column
cvint = []
df2['cvInt'] = ''
i = 0
while isinstance(df2.loc[i, 'cvCat'], str):
    if (df2.loc[i, 'cvCat'] == 'C') and (df2.loc[i+1, 'cvCat'] == 'V'):
        cvint.append((df2.loc[i, 'cvDur']) + (df2.loc[i+1, 'cvDur']))
        i = i + 2
    else:
        cvint.append(df2.loc[i, 'cvDur'])
        i = i + 1
    df2.loc[:, 'cvInt'] = Series(cvint)

#compute rhythm metrics
v_time = df2['Vowels'].sum()
c_time = df2['Cons'].sum()
sil_time = df2['Sil'].sum()
total_time = df2.loc[len(df2)-1, 'Stop']
total_speech_time = v_time + c_time

p_vowels = v_time / total_speech_time
p_sil = sil_time / total_time

std_vowels = df2['Vowels'].std()
std_cons = df2['Cons'].std()

varco_v = 100 * std_vowels / df2['Vowels'].mean()
APPENDIX B. PYTHON SCRIPT

```python
varco_c = 100 * std_cons/df2['Cons'].mean()

# rPVI for consonants
rPVI = 0
for i in range(len(cint)-1):
    curr = df2.loc[i, 'Cons']
    nxt = df2.loc[i+1, 'Cons']
    rPVI += np.abs(curr - nxt)
rPVI = rPVI / (len(cint) - 1.0)

# rPVI for CV intervals
rPVICV = 0
for i in range(len(cvint)-1):
    curr = df2.loc[i, 'cvInt']
    nxt = df2.loc[i+1, 'cvInt']
    rPVICV += np.abs(curr - nxt)
rPVICV = rPVICV / (len(cvint) - 1.0)

# nPVI for vowels
nPVI = 0
for i in range(len(vint)-1):
    curr = df2.loc[i, 'Vowels']
    nxt = df2.loc[i+1, 'Vowels']
    nPVI += np.abs((curr - nxt) / ((curr + nxt) / 2.0))
nPVI = 100.0 * (nPVI / (len(vint) - 1.0))

# calculate CV rate
cv_rate = (len(cvint))/total_speech_time

# create new df with stats
index = ['total_time', 'sil_time', 'p_sil', 'p_vowels', 'std_vowels',
         'std_cons', 'varco_v', 'varco_c', 'rPVI', 'rPVICV', 'nPVI', 'cv_rate']
data = [total_time, sil_time, p_sil, p_vowels, std_vowels, std_cons,
        varco_v, varco_c, rPVI, rPVICV, nPVI, cv_rate]
df3=pd.DataFrame(index=index, data=data)

# output stats df as csv
outFile2 = open(fname[:-4] + '.csv', 'w')
df3.to_csv(outFile2)
df3.close()
```

74
C.1 Entry for MDD in the DSM-5

This section contains the diagnostic criteria for Major Depressive Disorder in the DSM-5 (American Psychiatric Association 2013).

A. Five (or more) of the following symptoms have been present during the same 2-week period and represent a change from previous functioning; at least one of the symptoms is either (1) depressed mood or (2) loss of interest or pleasure. Note: do not include symptoms that are clearly attributable to another medical condition.

1. Depressed mood most of the day, nearly every day, as indicated by either subjective report (e.g. feels sad, empty, hopeless) or observation made by others (e.g. appears tearful). (Note: in children and adolescents, can be irritable mood.)

2. Markedly diminished interest or pleasure in all, or almost all, activities most of the day, nearly every day (as indicated by either subjective account or observation).

3. Significant weight loss when not dieting or weight gain (e.g. a change of more than 5% of body weight in a month), or decrease or increase in appetite nearly every day. (Note: In children, consider failure to make expected weight gain.)

4. Insomnia or hypersomnia nearly every day.

5. Psychomotor agitation or retardation nearly everyday (observable by others, not merely subjective feelings of restlessness or being slowed down).

6. Fatigue or loss of energy nearly every day.

7. Feelings of worthlessness or excessive or inappropriate guilt (which may be delusional) nearly every day (not merely self-reproach or guilt about being sick).
8. Diminished ability to think or concentrate, or indecisiveness, nearly every day (either by subjective account or as observed by others).

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

B. The symptoms cause clinically significant distress or impairment in social, occupational, or other important areas of functioning.

C. The episode is not attributable to the physiological effects of a substance or to another medical condition.

Note: Criteria A-C represent a major depressive episode

Note: Responses to a significant loss (e.g. bereavement, financial ruin, losses from a natural disaster, a serious medical illness or disability) may include the feelings of intense sadness, rumination about the loss, insomnia, poor appetite, and weight loss noted in Criterion A, which may resemble a depressive episode. Although such symptoms may be understandable or considered appropriate to the loss, the presence of a major depressive episode in addition to the normal response to a significant loss should also be carefully considered. This decision inevitably requires the exercise of clinical judgment based on the individual's history and the cultural norms for the expression of distress in the context of loss.

D. The occurrence of the major depressive episode is not better explained by schizoaffective disorder, schizophreniform disorder, delusional disorder, or other specified and unspecified schizophrenia spectrum and other psychotic disorders.

E. There has never been a manic episode or a hypomanic episode. Note: This exclusion does not apply if all of the manic-like or hypomanic-like episodes are substance-induced or are attributable to the physiological effects of another medical condition.
C.2 Quick Inventory of Depressive Symptomatology

An example of the self-report 16-item Quick Inventory of Depressive Symptomatology created by Rush et al. (2003) is included, along with its scoring procedure (University of Pittsburgh Epidemiology Data Center 2016).
# QUICK INVENTORY OF DEPRESSIVE SYMPTOMATOLOGY (SELF-REPORT)

**THIS SECTION FOR USE BY STUDY PERSONNEL ONLY.**

Questionnaire completed on visit date □ or specify date completed: __________________________

DD-Mon-YYYY

Only the patient (subject) should enter information onto this questionnaire.

**PLEASE CHECKMARK THE ONE RESPONSE TO EACH ITEM THAT IS MOST APPROPRIATE TO HOW YOU HAVE BEEN FEELING OVER THE PAST 7 DAYS.**

1. **Falling asleep:**
   - □ 0 I never took longer than 30 minutes to fall asleep.
   - □ 1 I took at least 30 minutes to fall asleep, less than half the time (3 days or less out of the past 7 days).
   - □ 2 I took at least 30 minutes to fall asleep, more than half the time (4 days or more out of the past 7 days).
   - □ 3 I took more than 60 minutes to fall asleep, more than half the time (4 days or more out of the past 7 days).

2. **Sleep during the night:**
   - □ 0 I didn’t wake up at night.
   - □ 1 I had a restless, light sleep, briefly waking up a few times each night.
   - □ 2 I woke up at least once a night, but I got back to sleep easily.
   - □ 3 I woke up more than once a night and stayed awake for 20 minutes or more, more than half the time (4 days or more out of the past 7 days).

3. **Waking up too early:**
   - □ 0 Most of the time, I woke up no more than 30 minutes before my scheduled time.
   - □ 1 More than half the time (4 days or more out of the past 7 days), I woke up more than 30 minutes before my scheduled time.
   - □ 2 I almost always woke up at least one hour or so before my scheduled time, but I got back to sleep eventually.
   - □ 3 I woke up at least one hour before my scheduled time, and couldn’t get back to sleep.

4. **Sleeping too much:**
   - □ 0 I slept no longer than 7-8 hours/night, without napping during the day.
   - □ 1 I slept no longer than 10 hours in a 24-hour period including naps.
   - □ 2 I slept no longer than 12 hours in a 24-hour period including naps.
   - □ 3 I slept longer than 12 hours in a 24-hour period including naps.

5. **Feeling sad:**
   - □ 0 I didn’t feel sad.
   - □ 1 I felt sad less than half the time (3 days or less out of the past 7 days).
   - □ 2 I felt sad more than half the time (4 days or more out of the past 7 days).
   - □ 3 I felt sad nearly all of the time.
<table>
<thead>
<tr>
<th>QUICK INVENTORY OF DEPRESSIVE SYMPTOMATOLOGY (SELF-REPORT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLEASE CHECKMARK THE ONE RESPONSE TO EACH ITEM THAT IS MOST APPROPRIATE TO HOW YOU HAVE BEEN FEELING OVER THE PAST 7 DAYS.</td>
</tr>
</tbody>
</table>

Please complete either 6 or 7 (not both)

6. Decreased appetite:
- 0 There was no change in my usual appetite.
- 1 I ate somewhat less often or smaller amounts of food than usual.
- 2 I ate much less than usual and only by forcing myself to eat.
- 3 I rarely ate within a 24-hour period, and only by really forcing myself to eat or when others persuaded me to eat.

7. Increased appetite:
- 0 There was no change in my usual appetite.
- 1 I felt a need to eat more frequently than usual.
- 2 I regularly ate more often and/or greater amounts of food than usual.
- 3 I felt driven to overeat both at mealtime and between meals.

Please complete either 8 or 9 (not both)

8. Decreased weight (within the last 14 days):
- 0 My weight has not changed.
- 1 I feel as if I’ve had a slight weight loss.
- 2 I’ve lost 2 pounds (about 1 kilo) or more.
- 3 I’ve lost 5 pounds (about 2 kilos) or more.

9. Increased weight (within the last 14 days):
- 0 My weight has not changed.
- 1 I feel as if I’ve had a slight weight gain.
- 2 I’ve gained 2 pounds (about 1 kilo) or more.
- 3 I’ve gained 5 pounds (about 2 kilos) or more.

10. Concentration/decision-making:
- 0 There was no change in my usual ability to concentrate or make decisions.
- 1 I occasionally felt indecisive or found that my attention wandered.
- 2 Most of the time, I found it hard to focus or to make decisions.
- 3 I couldn’t concentrate well enough to read or I couldn’t make even minor decisions.

11. Perception of myself:
- 0 I saw myself as equally worthwhile and deserving as other people.
- 1 I put the blame on myself more than usual.
- 2 For the most part, I believed that I caused problems for others.
- 3 I thought almost constantly about major and minor defects in myself.

12. Thoughts of my own death or suicide:
- 0 I didn’t think of suicide or death.
- 1 I felt that life was empty or wondered if it was worth living.
- 2 I thought of suicide or death several times for several minutes over the past 7 days.
- 3 I thought of suicide or death several times a day in some detail, or I made specific plans for suicide or actually tried to take my life.
## QUICK INVENTORY OF DEPRESSIVE SYMPTOMATOLOGY (SELF-REPORT)

PLEASE CHECKMARK THE ONE RESPONSE TO EACH ITEM THAT IS MOST APPROPRIATE TO HOW YOU HAVE BEEN FEELING OVER THE PAST 7 DAYS.

### 13. General interest:
- **0** There was no change from usual in how interested I was in other people or activities.
- **1** I noticed that I was less interested in other people or activities.
- **2** I found I had interest in only one or two of the activities I used to do.
- **3** I had virtually no interest in the activities I used to do.

### 14. Energy level:
- **0** There was no change in my usual level of energy.
- **1** I got tired more easily than usual.
- **2** I had to make a big effort to start or finish my usual daily activities (for example: shopping, homework, cooking or going to work).
- **3** I really couldn’t carry out most of my usual daily activities because I just didn’t have the energy.

### 15. Feeling more sluggish than usual:
- **0** I thought, spoke, and moved at my usual pace.
- **1** I found that my thinking was more sluggish than usual or my voice sounded dull or flat.
- **2** It took me several seconds to respond to most questions and I was sure my thinking was more sluggish than usual.
- **3** I was often unable to respond to questions without forcing myself.

### 16. Feeling restless (agitated, not relaxed, fidgety):
- **0** I didn’t feel restless.
- **1** I was often fidgety, wringing my hands, or needed to change my sitting position.
- **2** I had sudden urges to move about and was quite restless.
- **3** At times, I was unable to stay seated and needed to pace around.

---


---

<table>
<thead>
<tr>
<th>I confirm this information is accurate.</th>
<th>Patient’s/Subject’s initials:</th>
<th>Date:</th>
</tr>
</thead>
</table>

---
**QUICK INVENTORY OF DEPRESSIVE SYMPTOMATOLOGY (SCORE SHEET)**

NOTE: THIS SECTION IS TO BE COMPLETED BY THE STUDY PERSONNEL ONLY.

- ____ Enter the highest score on any 1 of the 4 sleep items (1-4)
- ____ Item 5
- ____ Enter the highest score on any 1 of the appetite/weight items (6-9)
- ____ Item 10
- ____ Item 11
- ____ Item 12
- ____ Item 13
- ____ Item 14
- ____ Enter the highest score on either of the 2 psychomotor items (15 and 16)
- ____ **Total Score (Range: 0-27)**


EPI0905.QIDSSR
C.3 Hamilton Rating Scale for Depression

An example of the Hamilton Rating Scale for Depression created by Hamilton (1960) is included (University of Massachusetts Medical School 2016).
## THE HAMILTON RATING SCALE FOR DEPRESSION

(to be administered by a health care professional)

**Patient's Name**

**Date of Assessment**

To rate the severity of depression in patients who are already diagnosed as depressed, administer this questionnaire. The higher the score, the more severe the depression.

**For each item, write the correct number on the line next to the item. (Only one response per item)**

<table>
<thead>
<tr>
<th>Number</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1.</strong></td>
<td><strong>DEPRESSED MOOD</strong> (Sadness, hopeless, helpless, worthless)</td>
</tr>
<tr>
<td></td>
<td>0= Absent</td>
</tr>
<tr>
<td></td>
<td>1= These feeling states indicated only on questioning</td>
</tr>
<tr>
<td></td>
<td>2= These feeling states spontaneously reported verbally</td>
</tr>
<tr>
<td></td>
<td>3= Communicates feeling states non-verbally—i.e., through facial expression, posture, voice, and tendency to weep</td>
</tr>
<tr>
<td></td>
<td>4= Patient reports VIRTUALLY ONLY these feeling states in his spontaneous verbal and non-verbal communication</td>
</tr>
<tr>
<td><strong>2.</strong></td>
<td><strong>FEELINGS OF GUILT</strong></td>
</tr>
<tr>
<td></td>
<td>0= Absent</td>
</tr>
<tr>
<td></td>
<td>1= Self reproach, feels he has let people down</td>
</tr>
<tr>
<td></td>
<td>2= Ideas of guilt or rumination over past errors or sinful deeds</td>
</tr>
<tr>
<td></td>
<td>3= Present illness is a punishment. Delusions of guilt</td>
</tr>
<tr>
<td></td>
<td>4= Hears accusatory or denunciatory voices and/or experiences threatening visual hallucinations</td>
</tr>
<tr>
<td><strong>3.</strong></td>
<td><strong>SUICIDE</strong></td>
</tr>
<tr>
<td></td>
<td>0= Absent</td>
</tr>
<tr>
<td></td>
<td>1= Feels life is not worth living</td>
</tr>
<tr>
<td></td>
<td>2= Wishes he were dead or any thoughts of possible death to self</td>
</tr>
<tr>
<td></td>
<td>3= Suicidal ideas or gesture</td>
</tr>
<tr>
<td></td>
<td>4= Attempts at suicide (any serious attempt rates 4)</td>
</tr>
<tr>
<td><strong>4.</strong></td>
<td><strong>INSOMNIA EARLY</strong></td>
</tr>
<tr>
<td></td>
<td>0= No difficulty falling asleep</td>
</tr>
<tr>
<td></td>
<td>1= Complains of occasional difficulty falling asleep—i.e., more than 1/2 hour</td>
</tr>
<tr>
<td></td>
<td>2= Complains of nightly difficulty falling asleep</td>
</tr>
<tr>
<td><strong>5.</strong></td>
<td><strong>INSOMNIA MIDDLE</strong></td>
</tr>
<tr>
<td></td>
<td>0= No difficulty</td>
</tr>
<tr>
<td></td>
<td>1= Patient complains of being restless and disturbed during the night</td>
</tr>
<tr>
<td></td>
<td>2= Waking during the night—any getting out of bed rates 2 (except for purposes of voiding)</td>
</tr>
</tbody>
</table>
6. **INSOMNIA LATE**
   0 = No difficulty
   1 = Waking in early hours of the morning but goes back to sleep
   2 = Unable to fall asleep again if he gets out of bed

7. **WORK AND ACTIVITIES**
   0 = No difficulty
   1 = Thoughts and feelings of incapacity, fatigue or weakness related to activities; work or hobbies
   2 = Loss of interest in activity; hobbies or work—either directly reported by patient, or indirect in listlessness, indecision and vacillation (feels he has to push self to work or activities)
   3 = Decrease in actual time spent in activities or decrease in productivity
   4 = Stopped working because of present illness

8. **RETARDATION: PSYCHOMOTOR** (Slowness of thought and speech; impaired ability to concentrate; decreased motor activity)
   0 = Normal speech and thought
   1 = Slight retardation at interview
   2 = Obvious retardation at interview
   3 = Interview difficult
   4 = Complete stupor

9. **AGITATION**
   0 = None
   1 = Fidgetiness
   2 = Playing with hands, hair, etc.
   3 = Moving about, can't sit still
   4 = Hand wringing, nail biting, hair-pulling, biting of lips

10. **ANXIETY (PSYCHOLOGICAL)**
    0 = No difficulty
    1 = Subjective tension and irritability
    2 = Worrying about minor matters
    3 = Apprehensive attitude apparent in face or speech
    4 = Fears expressed without questioning

11. **ANXIETY SOMATIC**: Physiological concomitants of anxiety, (i.e., effects of autonomic overactivity, "butterflies," indigestion, stomach cramps, belching, diarrhea, palpitations, hyperventilation, paresthesia, sweating, flushing, tremor, headache, urinary frequency). Avoid asking about possible medication side effects (i.e., dry mouth, constipation)
    0 = Absent
    1 = Mild
    2 = Moderate
    3 = Severe
    4 = Incapacitating
12. **SOMATIC SYMPTOMS (GASTROINTESTINAL)**

- **0= None**
- **1= Loss of appetite but eating without encouragement from others. Food intake about normal**
- **2= Difficulty eating without urging from others. Marked reduction of appetite and food intake**

13. **SOMATIC SYMPTOMS GENERAL**

- **0= None**
- **1= Heaviness in limbs, back or head. Backaches, headache, muscle aches. Loss of energy and fatigability**
- **2= Any clear-cut symptom rates 2**

14. **GENITAL SYMPTOMS** (Symptoms such as: loss of libido; impaired sexual performance; menstrual disturbances)

- **0= Absent**
- **1= Mild**
- **2= Severe**

15. **HYPOCHONDRIASIS**

- **0= Not present**
- **1= Self-absorption (bodily)**
- **2= Preoccupation with health**
- **3= Frequent complaints, requests for help, etc.**
- **4= Hypochondriacal delusions**

16. **LOSS OF WEIGHT**

- **A. When rating by history:**
  - **0= No weight loss**
  - **1= Probably weight loss associated with present illness**
  - **2= Definite (according to patient) weight loss**
  - **3= Not assessed**

17. **INSIGHT**

- **0= Acknowledges being depressed and ill**
- **1= Acknowledges illness but attributes cause to bad food, climate, overwork, virus, need for rest, etc.**
- **2= Denies being ill at all**

18. **DIURNAL VARIATION**

- **A. Note whether symptoms are worse in morning or evening. If NO diurnal variation, mark none**
  - **0= No variation**
  - **1= Worse in A.M.**
  - **2= Worse in P.M.**

- **B. When present, mark the severity of the variation. Mark “None” if NO variation**
  - **0= None**
  - **1= Mild**
  - **2= Severe**
19. DEPERSONALIZATION AND DEREALIZATION (Such as: Feelings of unreality; Nihilistic ideas)

   0 = Absent
   1 = Mild
   2 = Moderate
   3 = Severe
   4 = Incapacitating

20. PARANOID SYMPTOMS

   0 = None
   1 = Suspicious
   2 = Ideas of reference
   3 = Delusions of reference and persecution

21. OBSESSINAL AND COMPULSIVE SYMPTOMS

   0 = Absent
   1 = Mild
   2 = Severe

Total Score ______________


