ANALYSIS OF AFFECTIVE STATES FROM VOCAL ACOUSTICS IN ADULTS WITH APHASIA

A Dissertation
Presented to
The Academic Faculty

by

Stephanie Marie Gillespie

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering

Georgia Institute of Technology

May 2017

ANALYSIS OF AFFECTIVE STATES FROM VOCAL ACOUSTICS IN ADULTS WITH APHASIA

Approved by:

Dr. Elliot Moore, Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Mark Davenport
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Jacqueline Laures-Gore, Co-Advisor
Educational Psychology, Special Education, and Communication Disorders Department
Georgia State University

Dr. Pamela Bhatti
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Mark Clements
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Bruce Walker
School of Psychology, School of Interactive Computing
Georgia Institute of Technology

Date Approved: April 6th, 2017

ACKNOWLEDGEMENTS

Thank you to my family, for their unconditional love and support throughout my entire life and especially over the last five years. Thank you to my friends, for helping to keep me sane as I spent four years working on a single project with a seemingly never-ending set of goals and deadlines. I also thank my peers, Udit Gupta and Yash-Yee Logan, who were there for questions, support, edits, and advice.

Thank you to my committee members, Dr. Mark Davenport, Dr. Mark Clements, Dr. Pamela Bhatti, and Dr. Bruce Walker, for acknowledging the importance of this exploratory research that answered some questions, but left many more unanswered.

This work would not have been possible without Matthew Farina, previously at Georgia State University, for his assistance with data collection and labeling, and Scott Russel, of Grady Memorial Health, for referral of patients to the study. Special thanks are owed to Dr. Jacqueline Laures-Gore, who acted as a co-advisor and provided clinical guidance throughout the process. This dissertation would not exist if not for my advisor, Dr. Elliot Moore, due to his constant feedback, support, and never-ending belief in this research effort and my personal abilities.

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship, Grant No. DGE-1148903. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. Supported by the Emory-Georgia Institute of Technology Healthcare Innovation Program and the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number UL1TR000454. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Aphasia is a language disorder that may occur post-stroke that can impact quality of life and mental health due to its debilitating impact on the individuals who experience a sudden loss in their abilities to communicate. Stroke survivors who are living with aphasia are at risk to develop clinical stress and depression, but most diagnostic techniques for diagnosing stress and depression have not been adapted for those with aphasia or other language difficulties. Adapted assessments include self-assessment manikins which are mostly used to determine short-term emotional state, or caregiver assessments which ask questions related to behavioral changes the participant may have exhibited over the last month. These methods are subjective and do not utilize the methods that have progressed towards automatic stress and depression detection from speech analysis in the population without aphasia. It would be preferable for clinicians to have an objective tool that can analyze the vocal acoustics from speech of adults with aphasia to assist in their diagnostics of affective states.

This work focuses on analysis of a speech dataset from adults with aphasia collected by Georgia State University. The database participants present with a wide range of aphasia types, severities, ages, and other co-occurring motor disorders. Vocal acoustics were extracted and analyzed for their ability to detect depression, predict clinical stress and depression scores, and correlate to changes in short-term affective states. This exploratory work has investigated an area of affective analysis that the larger speech processing community has chosen to exclude due to the challenges of working with a population with a language disorder (and often co-occurring motor disorders as well).

Analysis of the vocal acoustics and literature from the speech processing community suggest the vocal acoustics may have been influenced by the presence of motor disorders in the participants. The same vocal acoustics which were studied for aphasia were extracted from databases containing speech from adults with dysarthria to determine to what extent the features could be used to predict the presence of dysarthria. The analysis of cross-database training and testing with respect to the vocal acoustics of dysarthria highlighted the challenges of combining datasets to overcome the often limited number of participants available for analysis, as well as the utility of some of the features to detect dysarthria.

This exploration of affect in aphasia through vocal acoustics has highlighted the need for signal processing research to explore the more challenging subsets of the population that are often excluded from affective analysis, even when they are at a higher risk for experiencing stress and depression. While more work is needed before a clinical tool will achieve the precision and recall necessary to be utilized in a clinical setting, this dissertation provides a foundation on which future studies will build.

CHAPTER I

INTRODUCTION

Each year, approximately 795,000 people will have a stroke, with 610,000 of these being new stroke victims [1]. Stroke is a leading cause of serious long-term disability in the United States [2], with 25-40% of stroke survivors developing aphasia [3, 4, 5]. Aphasia is a language disorder often resulting from stroke which can impair an individual's ability to read, write, comprehend auditory dialog, and express himself/herself verbally. An estimated two million people in the United States suffer from aphasia, with nearly 180,000 acquiring the disorder each year [3]. Due to the difficulty with language skills, individuals living with aphasia may be under considerable stress [6, 7] which can be a factor associated with increased risk of depression. However, work by Laures-Gore and DeFife indicated that most studies of post-stroke depression exclude adults living with aphasia due to comprehension and expression disabilities that many questionnaires cannot accommodate [8].

Accurate diagnosis of stress, depression, and determination of affective states in adults with aphasia is challenging due to the linguistic burden of many self-report measures, the potential for psychophysiological measures to be compromised by the neurological changes accompanying stroke, and problems associated with proxy-based questionnaires. The challenge of assessing stress, depression, and affective state in adults with aphasia is most obviously demonstrated by the exclusion of many adults with aphasia from post-stroke depression (PSD) studies due to their inability to complete standard depression questionnaires [9]. Often, post-stroke assessment of stress, depression, and affect in adults with aphasia is limited to proxy-based questionnaires or visual scale substitutions. Prior work has concluded with concerns that stroke

patients may have difficulty with visual analogue scales for mood [10]. Additionally, Bennett et al. found some visual analogue scales may be poor tools for screening low mood in patients with stroke, even though they could be useful over time for monitoring changes in mood [11]. Both prior studies excluded participants with aphasia.

Exclusion of adults with aphasia from research as well as the clinical implications of inaccurate diagnosis of mood disorders and affective states can have negative effects on post-stroke recovery, mental and physical health, and cognitive functioning [12]. Development of accurate and accessible techniques that avoid the pitfalls of conventional assessment methods would enable health professionals to more effectively treat mental health problems that develop during post-stroke recovery in adults with aphasia. The use of speech acoustics in diagnosing stress, depression, and affective states in adults with aphasia holds promise as a technique that permits avoidance of self-report questionnaires or behavioral observations and could potentially replace or augment conventional approaches to assessment. The overall goal of the work presented in this dissertation has been to investigate how vocal acoustic speech features and speech processing techniques could be used in objective diagnoses of perceived stress, clinical depression, and/or emotional state in the post-stroke population living with aphasia. Background material relating to aphasia, motor disorders, and speech analysis techniques is presented as a summary in Chapter 2, and details regarding the databases collected and used throughout the work are included in Chapter 3.

The first set of analyses on this work focused on the Aphasia Database. Chapter 4 highlights the preliminary investigation first which attempted classification of depression on the SADQ-10 score using a support-vector machine. Results indicated classification may not work well as a binary depression decision due to the ordinal nature of depression, so regression was utilized to predict SADQ-10 and PSS scores, presented in Chapter 5. Inconclusive results demonstrate the difficulty of working with limited participants with various clinical conditions. To overcome this challenge,

the changes within an individual's speech features were analyzed by correlating them to reported changes in acute affect. These results are presented in Chapter 6.

However, stroke survivors with aphasia may also have co-existing motor disorders, specifically apraxia of speech and dysarthria. While aphasia alone would not be expected to affect the vocal acoustics, motor disorders have been shown to change the acoustics due to difficulties in speech production. Chapter 7 reports on a cross-database experiment testing the ability to detect the presence of dysarthria using the same speech features as those analyzed in the Aphasia Database. This chapter illustrates how the same vocal acoustics can be used in the study of multiple different language disorders, motor disorders, and affective state analysis for various distinct purposes. Conclusions and future work relating to affective state analysis from vocal acoustics in adults with aphasia are presented in Chapter 8.

CHAPTER II

BACKGROUND

This work is multidisciplinary in the areas of electrical engineering and clinical speech-language pathology. The theory and methods originate from the digital signal processing field, specifically affect recognition in speech processing and machine learning. However, the application and impact of this work falls in the field of clinical communication disorders research. As such, it is important to understand both the methodology and the field from which the data has been collected. This literature review will broadly discuss the clinical speech disorders of interest to this work in Section 2.1, as well as the current techniques utilized in speech processing for affect recognition in Section 2.2.

2.1 Characteristics and Diagnosis of Aphasia, Dysarthria, and Apraxia

It is estimated that between 25% and 40% of stroke survivors will be diagnosed with aphasia [3, 4, 5], with estimates of 10-18% of survivors left with aphasia in the long term [4]. Post-stroke incidence rates of dysarthria range from 42-70% [13, 14], with an estimated 15% of stroke survivors diagnosed with both aphasia and dysarthria [14]. An understanding of speech and vocal acoustics impacted by aphasia (the language disorder of interest) and the often co-occurring motor disorders is essential for targeted machine-learning and analysis of results. As the post-stroke language disorder of interest, aphasia will be discussed in Subsection 2.1.1, while Subsection 2.1.2 will cover the motor disorder dysarthria and Subsection 2.1.3 will briefly discuss the motor disorder Apraxia of Speech (AOS).

2.1.1 Aphasia

Aphasia is a language disorder that can occur post-stroke. Specifically, aphasia occurs when there is a disruption to speech in regards to structure and linguistic rules. The impacts on language due to aphasia may include: phonology, syntax, semantics, and working memory [15]. Phonological changes may result in syllables or phonemes being swapped for another, while syntax difficulties may result in trouble understanding word orders. Semantic losses (also known as anomia) may result in disrupted word meaning while loss of working memory may result in difficulty understanding complex or longer sentences. The characteristics of the language produced are impacted differently depending on the type of aphasia present.

In general, aphasia can be grouped into two major categories. Fluent aphasia usually results in difficulty understanding spoken and written language. While sentence structure remains, there is a lack of meaning, and the speaker often has no idea of how they are speaking. Non-fluent aphasia usually results in difficulty communicating orally and with written words. While traditional noun/verb structures may be retained, sentences are difficult due to a lack of grammar. Benson attempted to distinguish between these two types of aphasia by considering the following: rate of speaking, prosody, articulation, phrase length, speech effort, pauses, press of speech, word choice, paraphasia, and verbal stereotypes (also known as recurring utterances) [16]. These features, in order of statistical importance in distinguishing between fluent and nonfluent aphasia as determined in a study by Kerschenstiner et al. were phrase length, pauses, prosody, rate of speaking, and effort [17].

A clinical diagnosis of aphasia is useful in generalizing the specific changes in speech caused by aphasia. Assessments of aphasia type are based on fluency, auditory comprehension, and repetition [15]. Fluent aphasics include Wernicke's aphasia, anomic aphasia, conduction aphasia, and transcortical sensory. As an example, Wernicke's aphasia is categorized by poor comprehension, resulting in spoken jargon or

nonsensical words and phrases. Nonfluent aphasics include Broca's aphasia, global aphasia, and transcortical motor aphasia. The characteristics of Broca's aphasia include pauses and effortful speech. Anomic aphasia is often the mildest aphasia (resulting in the fewest language problems) and is characterized by difficulties with word-finding tasks. It has been noted that most people with aphasia have anomic characteristics to at least some extent [15].

The Western Aphasia Battery-Revised (WAB) [18] is one aphasia diagnostic tool which determines the type of aphasia and an Aphasia Quotient (AQ), indicating how severe the aphasia is. The sections of the test include:

- Spontaneous speech
  - Conversational questions (e.g. How are you today?)
  - Picture description
- Auditory verbal comprehension
  - Yes/no questions
  - Auditory word recognition ("Point to the ___")
  - Sequential commands
- Repetition
  - Words and short phrases in a repeat-after-me manner
- Naming and word finding
  - Object naming ("What is this?", *points to an object*)
  - Word fluency (Name as many animals from memory in one minute)
  - Sentence completion (e.g. The grass is ___?)
  - Responsive speech (e.g. What do you write with?)

The participant's scores of each section are combined to determine the Aphasia Quotient. The aphasia type is determined from the values of each section and the range into which they fall. Possible aphasia classifications include global, Broca's, isolation, transcortical motor, Wernicke's, transcortical sensory, conduction, or anomic aphasia.

Most research on aphasia has focused on the linguistic changes in speech production. However, there is evidence that the vocal qualities of someone with aphasia may also change. In 1986, Ryalls discovered a difference in the standard deviations of the first and second formants in vowels spoken by male patients with aphasia in comparison to a non-aphasia control group [19]. While single tones spoken in the Thai language were found to have similar F0 contours between aphasics and non-aphasics, the F0 range was generally larger in aphasic and brain-damaged individuals than the non-aphasia controls [20].

Vukovic et al. performed a study of 60 Serbian-speaking aphasic patients, excluding those with dysarthria or apraxia [21]. Frequency alteration parameters, amplitude alteration parameters, and noise and tremor estimate parameters were considered in respect to the differences between the various types of aphasia patients as well as each group to a non-aphasic group. Results found that while all three aphasic populations varied from the control, Broca's aphasics and subcortical aphasics varied more so than Wernicke's aphasics. Specifically, "higher mean values for all parameters were present in participants with aphasia compared to speakers without stroke. Changes appeared greatest in association with subcortical aphasia, less in Broca's aphasia, and least in association with Wernicke's aphasia."

Temporal changes to speech by aphasia may also be of interest for analysis. Goldman-Eisler found that pauses in the general non-aphasic speech average 40-50% of the utterance time [22]. Due to this large percentage, it can be concluded that

pauses may be essential to the generation of spontaneous speech [23]. Quinting concluded that aphasic speech is directly impacted by both increased frequency and duration of pauses in speech as compared to non-aphasic speech. Specifically, he found that the length of unfilled pauses was measurably different between aphasic speech and nonaphasic speech.

From this section, it can be seen that aphasia has been studied from a prosodic standpoint in the clinical field for quite some time, but objective measures that extend beyond pitch, formants, and timing are often not considered by clinicians or researchers.

2.1.2 Dysarthria

Dysarthria is a term for a variety of speech motor disorders resulting from "paralysis, weakness, or incoordination of the speech musculature following damage to the central or peripheral nervous system" [15]. As such, dysarthria includes any disturbance of the basic components underlying speech production, including respiration, phonation, articulation, resonance, and prosody. Common speech errors due to dysarthria may include distortions such as slurred speech or mumbled speech. Dysarthria is often a consequence of a brain trauma (i.e., stroke) or a symptom of degenerative disorders (i.e., hypokinetic dysarthria for Parkinson's disease).

The traditional approach to diagnose dysarthria has clinicians perform perceptual judgments of speech to appraise the type and severity of the disorder, such as with the Frenchay Dysarthria Assessment [24]. The Frenchay Dysarthria Test- Second Edition (FDA-2) analyzes important parameters of the motor speech system to assist with neurological diagnosis of dysarthria and to guide treatment. Clinicians perform their analysis via subjective listening tests, analyzing a patient's reflexes, respiration, lips, jaw, palate, laryngeal, tongue, and intelligibility. From these sub-tests, a patient is given a severity ranging from 1-5, with one indicating mild and 5 indicating most

severe. There are some concerns that the clinical assessment procedure is costly, laborious, and prone to internal biases of the examiner [25]. The subjective nature of these tests has raised doubts about the reliability and validity of perceptual judgments by clinicians to consistently differentiate individual and coexistent speech disorders [26]. This motivates the design of a clinical tool that can objectively identify dysarthric speech from healthy speech and distinguish the type and severity of the dysarthria to assist with treatments and interventions.

Multiple research efforts to document the changes in speech and vocal acoustics due to dysarthria utilize analysis at the phoneme level in regards to spectral information of vowels, consonants, and transitions [27, 28, 29]. Tikofsky and Lehigh found that in vowel production, normal speakers tended to have a larger standard deviation in average F1 and F2 means than a group of dysarthric patients [28]. Lansford and Liss investigated both vowel-space metrics and F2 slope metrics to distinguish between dysarthric patients and the healthy control, as well as between those patients with various sub-types of dysarthria [29]. They were able to classify dysarthria from the control population at an accuracy of 84% using the mean dispersion of the front vowel space areas. It was found that average F2 slope was significant in distinguishing between the dysarthria subtypes- speakers diagnosed with hypokinetic dysarthria had a higher average F2 slope than those with ataxic or mixed-flaccid-spastic dysarthrias. Lansford suggests that the success of this feature is due to its representation of both spectral and temporal information. Kent et al. discussed the abnormal values found in the fundamental frequency variation, peak-amplitude variation, and smoothed amplitude perturbation quotient in dysarthric speech [30]. One challenge highlighted in the aforementioned studies is that features that may work well for one type of phoneme (consonants, fricatives, etc.) may not work well on others [31].

Other research efforts have focused on designing automatic speech recognition

(ASR) systems to help persons with dysarthria communicate [32, 33, 34]. Most speech processing work has analyzed dysarthria to broaden the applicability and real-world use of automatic speech recognition (ASR) systems. Shahamiri provides a comparison of several recent ASR models for dysarthric speech [34]. The UA-Speech dataset was collected with the goal of expanding ASR systems to a more diverse population including those with dysarthria [35], and has been used in multiple efforts for ASR [32, 34]. Other English dysarthric speech databases that exist are Nemours, Torgo, Whitaker and HomeService [36, 37, 38, 39]. A summary of the databases to which signal processing has been applied to dysarthria is available in Table 1. Work presented by Sriranjani et al. combined datasets of speech without dysarthria with the UA-Speech and Nemours Dataset to create larger datasets to train ASR systems, with results suggesting the incorporation of non-dysarthric data for models of dysarthric speech reduced performance [40]. While the goal of ASR with dysarthric speech has not focused on diagnosing or detecting dysarthria, Laaridh et al. highlight that one approach to automatic detection of dysarthria intelligibility could be the word transcript error rate based on automatic speech transcriptions systems [41] which often utilize ASR systems.

While ASR has been the primary focus of much dysarthria signal processing research, the task of automatically diagnosing and classifying the subtype and severity of dysarthria is less common. Carmichael worked towards a computerized application for detecting dysarthria subtypes by analysis of specific speech, articulatory, and phonation tasks [47]. However, the signal processing was applied to speech from tasks similar to a clinical dysarthria assessment as opposed to the more generic read or spontaneous speech. Mujumdar created a dysarthria subtype tree-based classification algorithm based on global prosodic and spectral statistics, achieving 90% accuracy across 35 moderately-severe dysarthric patients when selecting dysarthria subtypes that are known to exhibit distinct speech changes and patterns from each

Table 1: Selected dysarthria speech databases available for research purposes to the signal processing community

Database Name	Date Pub.	Etiology and Dysarthria Type	Collection Details	Cost/Access	Original Purpose	Selected Publications
Aronson Tapes	1993	unknown	Cassette recordings of 35 subjects with moderate severity and 1 control reading a passage	unsure	teaching tool for speech pathologists	Source: [42] [43]
Whitaker Dataset	1993	Cerebral Palsy	6 individuals spoke repetitions of isolated words and Also includes 1 control	free	ASR	Source: [38]
Nemours Dataset	1996	unknown	11 male speakers of varying degrees of dysarthria speaking 74 nonsense sentences and two read paragraphs, includes 1 control	$100 at time of publication	unknown	Source: [36], [40]
UA-Speech Dataset	2008	Cerebral Palsy Spastic,Flaccid	15 participants and 13 controls recorded isolated words	Free	ASR	Source: [35], [44, 25]
TORGO Dataset	2012	Cerebral Palsy or ALS	8 participants and 7 controls read syllabic repetitions, vowels, short words, passage, and picture description. Also includes video facial trackers and EMA kinematics	$1200 for nonmember, free if LDC member	ASR	Source: [37], [45, 46]

other [43]. DeMino automatically classified dysarthria severity into mild, moderate, or severe in 39 male and female participants [48]. Current research efforts to use full speech sentences for classification are limited and have not conclusively solved the automatic diagnosis of dysarthria subtype.

An important consideration of the work mentioned in this section is that often the dysarthria studied is that which is caused by progressive neuromuscular disorders including Parkinson's disease and amyotrophic lateral sclerosis (ALS) [15]. There is relatively little literature relating specifically to post-stroke dysarthria assessment and/or treatment.

2.1.3 Apraxia of Speech

Apraxia is defined to be "a disturbance of movement or action not due to neuromuscular innervation disruption, ... but instead to a hypothesized disruption of motor programming" [49]. Apraxia of speech (AOS) has been defined as

> "an articulatory disorder resulting from impairment, as a result of brain damage, of the capacity to program the positioning of speech musculature and the sequencing of muscle movements for volitional production of phonemes. The speech musculature does not show significant weakness, slowness, or incoordination when used for reflex and automatic acts. Prosodic alterations may be associated with the articulatory problems, perhaps in compensation for it." Defined by Darley, 1968 [50].

Apraxia is usually divided into various categories, including apraxia of speech, oral or facial apraxia, and limb apraxia. While specific definitions and localization areas related to speech programming are still under debate, apraxia of speech is generally described as an articulation- and prosody of speech- disorder. Due to these motor-difficulties, the speech produced by those with apraxia may change. The rate of speech due to apraxia tends to be slower than that of normal speech, and additional pauses

at the start of a word or in between syllables may be added [15][49]. Additionally, Cherney noted that syllables may be equally-stressed within an utterance and normal variations of pitch and loudness may be limited [15].

The Apraxia Battery for Adults- Second edition (ABA-2) determines the presence and severity of apraxia in adults [51]. Segments of the test determine diadochokinetic rate, increasing word length, limb apraxia, oral apraxia, utterance timing, ability of repetition, and articulation measures. Labels from each subsection result are 'none', 'mild', 'moderate', or 'severe'.

Even though Apraxia of Speech and aphasia share some common speech errors, they are unique disorders [52]. As one example, AOS and conduction aphasia have similar sound-level errors, but have a different underlying reason for the errors. Conduction aphasia sound errors relate to the underlying deficit in the selection of phonemes for speech. Apraxia sound errors are the correct phonemes, but are disrupted due to motor programming flaws.

2.2 Affect Recognition and Speech Processing

This section of the literature review will briefly summarize the field of speech processing as it relates to affect recognition. Subsection 2.2.1 will discuss affect recognition, specifically for general emotions, stress, and depression. Subsection 2.2.2 will look at the small amount of work in affect recognition from speech in persons with language disorders.

2.2.1 Affect Recognition from Speech

Affective computing relates to emotion recognition, analysis, and synthesis [53]. When using speech, affective speech processing often attempts to classify or detect various emotions (e.g., joy, anger, sadness, surprise) or emotional states (e.g., stress, depression, frustration) from analysis of the vocal acoustics from the speaker. In this work, clinical depression and perceived stress are of particular interest due to their clinical

implications as they relate to post-stroke mental health.

The detection of depression in the voice is directly related to a large foundation of research in the study of affect (emotion, stress) in the voice. Several literature surveys over the years [54, 55, 56, 57] have provided overviews and updates on work related to the creation and analysis of speech emotion databases and classification of emotion. Emotions collected for study are acted, elicited via a task, or natural (e.g., taken from other media with no original intent of being used in emotion recognition). Zeng et al. mentioned that while automatic detection of the six basic emotions has achieved high success rates when using controlled audio or visual displays, detection from less constrained settings such as spontaneous unscripted or unprompted speech still appears challenging [57]. Features range from prosodic, spectral, glottal, and non-linguistic vocal features (e.g., laughter, sighs, yawns). Additionally, various machine-learning algorithms are utilized ranging from k-nearest neighbor (k-NN) to Support Vector Machines (SVM) and decision trees, the model choice depending on the affective task of interest and available data. The main conclusion drawn by the various studies and other recent work is the lack of a one-size-fits-all feature set and model to handle all vocal emotion recognition tasks.

One specific area of affect recognition is the detection of stress, with multiple potential uses including medical diagnosis, interrogation analysis, and call center customer service. Zhou et al. found that Teager Energy Operator (TEO) features increased classification performance over pitch and MFCC features [58]. However, much of the work on stress detection utilizes speech under workload-simulated stress (e.g., [59]) or short-term stress (e.g., the SUSAS database [60]) instead of clinical stress. As such, there is a lack of work in the area of speech processing focusing on the detection of long-term stress in vocal acoustics.

Detection of depression through voice analysis is important for clinical research and could reduce the incidence of suicide. While not a new topic of interest in the

research community, i.e., [61, 62], depression of detection using vocal acoustics is still being studied due to the complications of detecting a long-term clinical diagnosis from voice acoustics that often vary in their short-term emotional state as well. The Audio/Visual Emotion Challenge and Workshop (AVEC) series has provided an avenue for research groups to compete for the best depression and emotion-recognition classification accuracies when using a standardized dataset [63, 64]. A recent review article by Cummins et al. [65] summarized speech analysis in depression and suicide risk over the last 10 years, including meta-analysis on vocal features as they relate to diagnosis and classification of depression. Similar to the features for emotion recognition, depression detection often utilizes prosodic, glottal, formant, and spectral features. Cummins et al. mention SVM and Gaussian Mixture Models (GMM) as two of the most popular prediction techniques for depression presence and severity [65], and concludes with a call to action for better research collaboration and further research into various demographic factors that lead to variability in depressed voices. This work addresses the call as aphasia is often an exclusion factor for speech analysis studies, even though adults with aphasia are more-likely to be depressed than the non-aphasia population.

2.2.2 Affect Recognition in Speech with Language Disorders

Section 2.1 introduced the language disorder aphasia and the motor disorders dysarthria and apraxia of speech. While Subsection 2.2.1 highlights the extensive body of literature for the recognition of stress and depression in persons without aphasia, there is an analogous lack of such research for persons with aphasia. While speech processing techniques have been used to determine aphasia intelligibility [66] and aphasia rehabilitation [67], the author is unaware of any prior research besides that mentioned in Chapters 4, 5, and 6 as work focused on affective speech processing in persons with aphasia. This lack of research is not surprising considering that aphasia affects the

production and/or language coherency of persons who suffer from it, and this population is often excluded from other research studies due to their language difficulties. Given the success of detecting stress, depression, and affect in other populations via speech analysis, exploration of this type of analysis in adults with aphasia could eventually lead to an accessible and accurate technique for identifying mood disorders in adults with aphasia.

CHAPTER III

DATASETS AND ACOUSTIC FEATURES

The following chapter will discuss the databases collected and used in this study, as well as the feature extraction techniques that were used.

3.1 Databases

Multiple databases were utilized in this work for a variety of purposes. The main goal of this dissertation has been to analyze affect from speech acoustics of adults with aphasia. In this work, the main database will be referred to as the 'Aphasia Database' which contains speech from adults with aphasia and is summarized in Section 3.1.1. While the participants presented with aphasia and various types of affective states (i.e., stress, depression, current emotional state), they also presented with other clinical conditions including apraxia of speech and dysarthria. The second database analyzed in this work is the Atlanta Motor Speech Disorder Corpus (AMSDC) which contains speech from adults with dysarthria. The AMSDC is summarized in Section 3.1.2. However, since the AMSDC has not been analyzed from a signal processing perspective, baseline experiments were conducted on the more-established UA-Speech dataset, which is detailed in Section 3.1.3.

3.1.1 Aphasia Database

Speech from 26 adults who were at least one-month post-onset of stroke was collected at the Aphasia and Motor Speech Disorders Laboratory at Georgia State University over a period of approximately one year from spring 2014 to summer 2015. Approval from the institutional review boards at Georgia State University and Georgia Institute of Technology was obtained prior to initiation of the study. Grady Memorial Hospital

approved the research through the Grady Research Oversight Committee. Participants in the study exhibited Broca's, Wernicke's, Conduction, and Anomic aphasia as determined by the Western Aphasia Battery (WAB) [18]. The WAB also assigns an Aphasia Quotient (AQ) that assesses the severity of the subject's aphasia. The range of values for the AQ is from 0-100 (most to least severe) with a score higher than 93.8 within normal limits indicating no aphasia. A total of two participants scored as normal on the WAB, a total of five participants had technical difficulties during the recording process, and one participant was removed after no proof of stroke could be determined. Table 2 shows selected demographic and clinical information for the participants with confirmed history of stroke and complete audio recordings.

Participants were recorded during their series of tasks as a part of the Western Aphasia Battery [18] as detailed in Section 2.1.1. Speech was recorded with an AKG C520 headset condenser microphone and sampled at 16 kHz. Approximately 55 responses per participant were segmented, ranging from single word answers to an extended description. An additional two picture descriptions, the cookie theft picture from the Boston Diagnostic Aphasia Exam [68] and the Cat in Tree sketch [69], were included in an attempt to elicit additional spontaneous speech. Longer responses, including those of the picture descriptions, were segmented into individual utterances based on the completion of an idea. In total, each participant had at least 3.5 minutes of responses recorded after segmentation and approximately 75 utterances.

Two proxy-based questionnaires specifically for identification of depressive symptoms in adults with aphasia exist: the Stroke Aphasia Depression Questionnaire (SADQ-10) [70] and the Aphasia Depression Rating Scale (ADRS) [71]. The SADQ-10 is a ten-question survey requiring caregivers to assess the frequency of specific behaviors of the participant, ranging from never (0) to always (3) [70]. In a subsequent study among patients without aphasia, a threshold of 14/30 was found to indicate clinically significant symptoms of depression with 70% sensitivity and 77%

Table 2: Clinical and demographic information for participants with confirmed stroke history and complete recordings

ID	Gender	Race	Age	Aphasia Type	Aphasia Quotient	Dysarthria Severity	Apraxia Of Speech Severity	SADQ-10 Score	PSS	Stress Scale-Pre	Stress Scale-Post	SAM Arousal Pre	SAM Arousal Post	SAM Valance Pre	SAM Valance Post
6	M	W	61	Anomic	87.4	none	mild	13	28	1	1	3	1	3	2
7	M	B	49	Anomic	82.1	mild	mild	10	23	2	2	1	1	2	2
8	F	B	46	Broca's	59.6	none	moderate-severe	19	26	1	1	1	1	1	1
9	M	W	70	Anomic	83.2	none	mild	8	14	1	1	1	1	1	1
10	M	B	32	Anomic	99.4	none	none	9	18	1	1	1	1	2	1
11	M	B	55	Anomic	78.0	moderate	mild-moderate	19	36	7	4	3	3	5	5
12	M	B	52	Broca's	58.3	mild	moderate-severe	11	30	3	5	1	2	1	2
13	M	B	57	Anomic	87.4	mild	none	16	28	2	5	3	3	3	2
14	M	B	67	Anomic	98.4	mild	none	22	38	4	3	3	3	2	2
15	F	B	51	Wernicke's	41.0	none	moderate-severe	12	33	1	7	1	5	1	5
16	F	W	39	Anomic	93.2	mild	mild	16	23	3	4	2	3	2	3
17	F	B	33	Anomic	92.4	mild	none	19	35	1	1	1	1	1	1
18	M	B	63	Anomic	83.2	mild	mild	20	24	4	1	2	1	1	1
19	F	B	52	Anomic	88.3	mild	mild	25	40	5	1	4	1	4	4
20	F	W	36	Conduction	54.6	none	moderate	12	22	3	3	3	4	3	3
21	M	W	49	Anomic	83.3	none	mild	16	24	1	1	1	1	2	2
23	M	W	31	Broca's	31.9	mild	mild-moderate	6	-	1	1	1	1	1	1
24	F	B	64	Broca's	68.5	mild	moderate	11	40	5	5	3	2	3	3
25	M	B	40	Wernicke's	75.8	mild	mild-moderate	14	32	1	2	1	1	1	5
26	M	B	25	Anomic	92.3	none	none	-	23	4	3	3	2	3	3

specificity [72]. SADQ-10 scores were received for 25 participants, with participant SADQ-10 scores ranging from 6-25.

Caregivers also completed the Aphasia Depression Rating Scale (ADRS) [71]. The ADRS is a nine-item questionnaire that rates external symptoms of depression such as insomnia, weight loss, outward signs of anxiety, and fatigability. The ADRS is scored by adding item scores into a combined score, with a higher score indicating more depressive symptoms. The ADRS authors established a threshold of 9/30 to suggest the presence of depression (83% sensitivity and 71% specificity). Based on work by Laures-Gore et al., ADRS and SADQ-10 scores were highly correlated ($r = 0.708$, $p < 0.001$) [73]. As such, SADQ-10 scores were selected for the analysis of depression in this work over those of the ADRS due to the higher completion rate.

The Perceived Stress Scale (PSS) is a fourteen-question assessment to determine the degree to which situations in an individual's life were considered stressful by that individual [74]. For each item, a user can respond with never (item score of 0) to very often (4), for a total of 56 points. While the PSS has been used in studies on those with aphasia [8], it was not designed specifically for aphasic populations. Therefore, administration of the PSS generally requires the assistance of a caregiver or interviewer who reads the questionnaire aloud as the person with aphasia selects responses. Participant PSS scores ranged from 14-40 with higher scores indicating higher levels of perceived stress. The PSS was not designed with severity thresholds or categories, nor can the authors find any published study validating proposed categories. However, some sources that publish the PSS-10 (a 10-question version of the test with max score of 40) advertise the severities as low stress (0-13), moderate stress (14-26), and high stress (27-40) [75, 76].

Acute assessments of affective state were collected at the start and end of the testing session. Participants were asked to complete valance and arousal Self-Assessment

Manikins [77], creating pre- and post- SAM scores. The non-verbal assessment consists of two sets of line-drawn figures expressing different emotional scales for a total of five figures in each set. While the original testing procedure for SAMs created output scores on a 9-point scale by allowing participants to also choose an emotional state between the pictures, the participants in this study were limited to selecting the picture that best represented their mood, limiting the numeric outputs to 1-5. An additional acute assessment was completed before and after the recording process. Identified in this work as the Stress-Scale, participants were asked to identify their emotion on a pictorial scale ranging from a sketched face shown smiling to a face frowning with a wrinkled brow [78]. The assessment output was a number between one and seven, identified by the authors as calm and stressed respectively. Both the SAM and Stress-Scale were used in this work as acute measures of affect as they allow feedback directly from the participants with linguistic challenges due to aphasia.

3.1.2 Atlanta Motor Speech Disorders Corpus (AMSDC)

The Atlanta Motor Speech Disorders Corpus (AMSDC) contains speech recordings of 99 adults local to the South-Eastern US with acquired neurogenic disorders that resulted in a motor speech disorder [79]. Participants presented with aprosodia, dysarthria, and apraxia of speech. An important distinction in this corpus is the emphasis on regional dialects unique to the southeastern US, especially those of African-American English. Audio recordings were completed in private or semi-private rooms of the Grady Memorial Health System using an AKG C520 condenser microphone with a behind-the-neck band. All participants completed recordings of conversational speech, oral readings of 'The Caterpillar'[80] and 'The Grandfather'[81] passages, sentences with contrasting stress emphasis on specific words, sentences with material sensitive to dysarthria, and affective sentences.

Table 3: Clinical and demographic information summaries for the 57 participants from the Atlanta Motor Speech Disorders Corpus included in analysis in this dissertation

	Females	Males
Number of Participants	22	35
Age Range	38 - 76	30 - 79
Number with Spastic Dysarthria	4	3
Number with Flaccid Dysarthria	4	9
Number with Mixed Dysarthria	11	11
Number with Other Dysarthria Subtypes	3	12
Very Low Participants	1	1
Low Participants	2	2
Middle Participants	2	7
High Participants	17	25

A subset of 57 participants who presented with dysarthria were selected for analysis in this work. For each participant, clinical and demographic information was available including dysarthria type and dysarthria intelligibility. Intelligibility was determined by a clinician and 4 graduate speech-language pathology interns based on intelligibility of the read paragraphs. For use in this work, participants' intelligibility was categorized into four quartiles defined as very low (0-25%), low (26-50%), middle (51-75%), and high (76-100%). Table 3 summarizes the clinical and demographic information of the AMSDC participant subset included in the work presented in this dissertation

3.1.3 Universal-Access Research Dataset (UA-Speech)

Originally collected by the University of Illinois, the Universal-Access Research (UA-Speech) dataset contains speech recordings from 15 speakers (4 female, 11 male) of mostly spastic dysarthria due to cerebral palsy [35]. Additionally, 13 controls without dysarthria (4 female, 9 male) were included for a total of 28 participants. Subjects were asked to read single isolated words shown on a laptop screen, with

Table 4: Clinical and demographic information summaries for the 15 participants and 13 controls from the UA-Speech Database included in analysis in this dissertation

	Females	Males	Female Controls	Male Controls
Number of Participants	4	11	4	9
Age Range	18 - 51	18 - 58	age matched	age matched
Number with Spastic Dysarthria	3	8	-	-
Number with Other Dysarthria	1	3	-	-
Very Low Participants	1	3	-	-
Low Participants	1	2	-	-
Middle Participants	1	2	-	-
High Participants	1	4	-	-

words representing the 10 digits, 26 radio alphabet letters, computer commands, common words from the Brown corpus of written English, and uncommon words from children's novels selected to maximize phone-sequence diversity, for a total of 765 isolated words, 455 of which were distinct. An eight-microphone array was used to collect the speech at 48 kHz, and processed to create seven channels. In this work, one channel was randomly selected per participant to be included in the analysis (as opposed to all seven channels) to reduce the number of repetitions within each individual's speech.

Speech intelligibility was rated by five native speakers without either transcription experience or experience working with individuals with speech disorders. A subset of the data was used, including 225 words per participant. Each listener was asked to transcribe the word as well as their confidence in their transcription. The accuracy of the transcriptions across all 5 participants and all 225 words was calculated to create an intelligibility measure for each individual ranging from 0 (worst) to 100 (best). Participants were categorized into four quartiles defined as very low (0-25%), low (26-50%), middle (51-75%), and high (76-100%). Table 4 summarizes the demographic information for the participants included in this study from the UA-Speech dataset.

3.2 Speech Features Extracted

In this work, a variety of prosodic, spectral, TEO, and glottal features were extracted from the various databases using Matlab. For each feature extracted, various low level descriptors (e.g. mean, inter-quartile range, skew) were calculated based on those described in the openSMILE software [82]. The features to which the low level descriptors (LLDs) were applied are described below.

Cepstral Peak Prominence (CPP) was computed by the VoiceSauce application using Hillenbrand's formula [83, 84]. The cepstral domain is a mathematical transformation of the signal from the frequency domain of the signal. Cepstral Peak Prominence is a measure of cepstral peak amplitude normalized by the overall amplitude. CPP has been found to be indicative of breathy speech, the overall periodicity of speech, and dysphonia severity [84, 85, 86].

Four different Harmonic-to-Noise Ratios (HNR) were computed by the VoiceSauce application using de Krom's method [83, 87]. HNR is defined as the level difference between the original spectrum and the noise spectrum of a signal within a specified frequency band. Previous studies have found HNR to be useful in identifying depression and suicide risk [65] as well as stress [88].

14 mel-frequency cepstral coefficients (MFCCs) and their deltas were computed by the VoiceBox toolbox [89]. Cepstral coefficients have been shown to be related to the human perception of sound, and the mel-scale more closely relates to the nonlinear frequency scale heard by humans. MFCCs are one of the most prevalent feature types for speech affect analysis, with previous research using MFCCs for detection of hypokinetic dysarthria as a symptom of Parkinson's Disease [90, 91], long-term stress [92], and depression [65, 93]. As such, they have become established as the basic feature vector for most acoustic speech recognition problems [94].

Pitch, based on Sun's Subharmonic-to-harmonic ratio [95] and jitter were also calculated. Pitch is the perceptual measure most closely related to frequency, while

jitter is the short-term, cycle-to-cycle perturbation in the fundamental frequency. Pitch and jitter have both been shown to describe the pathological characteristics of the speech, particularly in depression [65], emotional state [96, 97], and stress [92].

Eight Line Spectral Frequencies (LSF) were derived from the Linear Predictive Coefficients (LPCs). The LPCs are the mathematical coefficients that attempt to recursively identify a speech signal. The LSFs are the frequencies of the zeros to the equations defining the line-spectral pair parameters, a mathematical attempt to model the acoustics of speech [94]. While traditionally used for speech coding or analysis, LSFs have more recently been used in affect recognition [98].

Root Mean Square-Energy (RMS-Energy) is the square root of the mean of the squared value of the energy in a frame within a signal. RMS-Energy has detected valance and arousal [99] and determined depression from controls groups [100].

The Teager Energy Operator (TEO) is a nonlinear equation that was created to take advantage of short-term information within a signal from an energy point-of-view. TEO features were originally developed to detect stress [58], but have more recently been used to detect depression [101]. Specific TEO feature extracted in this work include: amplitude modulation, frequency modulation, 16 critical band areas [58], RMS-energy, and Log-energy.

The glottal waveform is an estimation of the volume velocity profile of the air as it passes through the glottis and speech folds during speech production. The features extracted relate to the timing, ratio, amplitude shimmer, and spectrum characteristics of the estimated waveform. These features have been shown to be useful in the detection of depression [102, 103, 104]. Specifically, the glottal parameters include: H1-H2 as the difference between the first and second glottal formants in decibels [105], Harmonic Richness Factor (HRF) as a measure of the ratio of higher harmonics to the first harmonic [106], Parabolic Spectrum Parameters (PSP) which matches a second order polynomial to the flow spectrum computed over a signal glottal cycle

[107], and a set of time parameter estimates to analyze various speeds and ratios of the opening and closing of the glottis [108].

CHAPTER IV

CLASSIFICATION OF DEPRESSION FROM SPEECH IN ADULTS WITH APHASIA

One of the main interests in collected the Aphasia Database was to investigate objective acoustic features relating to the detection of clinical depression. As the literature in Chapter 2 summarized, depression classification has been explored through speech analysis, but adults with aphasia have been excluded due to their language difficulties, even though they experience higher rates of stress and depression than the standard population. To investigate this research question, traditional acoustic vocal features were extracted and a linear support-vector machine (SVM) was used to create a binary classifier to assign the label of depressed/not-depressed. This chapter will highlight the preliminary depression classification results from the Aphasia Database, previously published by Gillespie et al. [109],

4.1 Methodology

The Aphasia Database included depression scores from both the SADQ-10 and the ADRS. However, the SADQ-10 was selected for analysis in this dissertation due to its higher completion rate. The community Stroke Aphasia Depression Questionnaire-10 (SADQ-10) was developed to assess depressed mood in individuals with aphasia [70]. As one of the few depression scales available for adults with aphasia, the SADQ-10 was used to determine the potential presence of depression. A participant's score of greater than 14 is assigned a label of high depressive symptoms [72]. In this proposal, the term 'depressed' is used as shorthand to indicate high depressive symptoms while 'not depressed' indicates low depressive symptoms.

Table 5: Participants' depression label, and aphasia type (aphasia quotient (AQ)) for participants chosen in depression classification analysis.

Participant Number	Gender	SADQ Score	Label (Depressed if SADQ-10 >14)	Aphasia Type (AQ)
15	Female	12	Not Depressed	Wernicke's (41)
20	Female	12	Not Depressed	Conduction (54.6)
24	Female	11	Not Depressed	Broca's (68.5)
8	Female	19	Depressed	Broca's (59.6)
19	Female	25	Depressed	Anomic (88.3)
5	Female	16	Depressed	Anomic (92.2)
25	Male	14	Not Depressed	Wernicke's (87.4)
7	Male	10	Not Depressed	Anomic (82.1)
9	Male	8	Not Depressed	Anomic (83.2)
6	Male	13	Not Depressed	Anomic (87.4)
11	Male	19	Depressed	Anomic (78.0)
18	Male	20	Depressed	Anomic (83.2)
21	Male	16	Depressed	Anomic (83.3)
13	Male	16	Depressed	Anomic (87.4)

Table 5 shows the demographics for the participants chosen for the classification task of depression based on SADQ-10 score. The 14 participants were chosen to be approximately balanced in gender and depression label, while still including as many participants as possible.

For the analysis, 33 responses were selected per person, with phrases selected over single-word answers based on initial analysis comparing classification accuracies of models trained on words and models trained on phrases. Prosodic and spectral features were extracted based on the features used in the INTERSPEECH 2009 emotion challenge [110] with some additional common features and statistics from openSMILE [82]. Specifically, low-level descriptors (LLD) statistics for the prosodic and spectral features detailed in Section 3.2 were calculated yielding a total of 874 features.

Features and their statistics were extracted in MATLAB for each utterance of each participant and Weka [111] was used for building the classifier using the SMO-SVM (Sequential Minimal Optimization-Support Vector Machine). Feature selection

on the full data set as well as each individual feature type was performed using 5-fold cross validation in Weka with the Correlation-Based Feature Subset selection algorithm [112] and the Best-First search method. Only those features that were selected by at least 3 of the 5 folds were chosen to be used to train and test the feature-subset classifiers. Restricted by the small dataset, a leave-one-out approach was used in which a training model was built on 13 participants and tested on the excluded participant, yielding 14 separate classification tasks per feature subset. Analyses were conducted based on all 874 features and each reduced subset of features grouped by the following categories: Pitch+Jitter, RMS-Energy, Harmonic Noise Ratio (HNR), Cepstral Peak Prominence (CPP), MFCC+deltas (MFCC + Δ), and LSF+deltas (LSF + Δ).

4.2 Results and Discussion

Table 6 shows the precision (i.e., percentage correctly marked as depression out of all observations marked as depression by the classifier), recall (i.e., percentage correctly marked as depression out of all possible true labels of depression), and average accuracy for both the original full feature set and each set of reduced features after the feature selection took place. Precision and recall balance each other as a high recall and low precision would indicate a tendency to correctly classify more of the true labels of depression at the cost of misdiagnosing a greater number of participants. A high precision and low recall would indicate a tendency for the classifier to miss more true diagnoses of depression while being more certain of the samples marked as depression. The values in Table 6 are based on assigning the appropriate label of depressed/not-depressed based on an individual's SADQ-10 score across their 33 utterances. A total of 462 utterances were classified across all 14 participants.

Based on the results in Table 6, the Cepstral Peak Prominence feature subset classified the best overall considering recall, precision, and accuracy. Only RMS-

Table 6: Depression classification results by feature subtype in assigning the correct depression label to each sentence, © IEEE 2016.

Features (number of features)	Avg. Recall	Avg. Precision	Avg. Accuracy (standard dev.)
All (874)	0.359	0.411	0.422 (0.264)
Reduced (41)	0.459	0.447	0.446 (0.325)
Pitch + Jitter (7)	0.394	0.399	0.400 (0.303)
RMS-Energy (8)	0.814	0.487	0.478 (0.478)
HNR (10)	0.545	0.472	0.468 (0.311)
CPP (6)	0.563	0.634	0.619 (0.190)
MFCC+Δ (19)	0.432	0.588	0.502 (0.349)
LSF+Δ (20)	0.308	0.286	0.374 (0.246)

Energy had a higher recall, but also had a lower precision and accuracy, indicating a tendency to assign the label of depression more often resulting in misdiagnoses. Cepstral Peak Prominence has been found to be indicative of breathy voices and has been extended to an evaluation of the overall periodicity of the speech [84, 86]. While the exact quantities that CPP measures are unknown, it is suggested that the CPP integrates aperiodicity and other waveform features [113]. Jitter, a similar feature that measures the noise of the signal's pitch, has been found to be statistically significant in classifying between suicidal and not-depressed participants, as well as between depressed and control participants [62, 100]. Understanding the jitter and CPP of depressed aphasic speech and not-depressed aphasic speech will need to be continued in further detail to determine if this is unique to aphasic speech exhibiting depressive symptoms or if the motor disorders of dysarthria and apraxia are also contributing to these differences found in these features.

MFCC and Δs were the next highest performing group of features overall regarding accuracy and precision, while RMS-Energy had the highest recall of any feature subset. The worst performance was from the LLDs of the LSF+Δ group which exhibited a classification accuracy well-below chance (approximately 0.5). While all of the features included above have been used in the past to assess depression, they are also

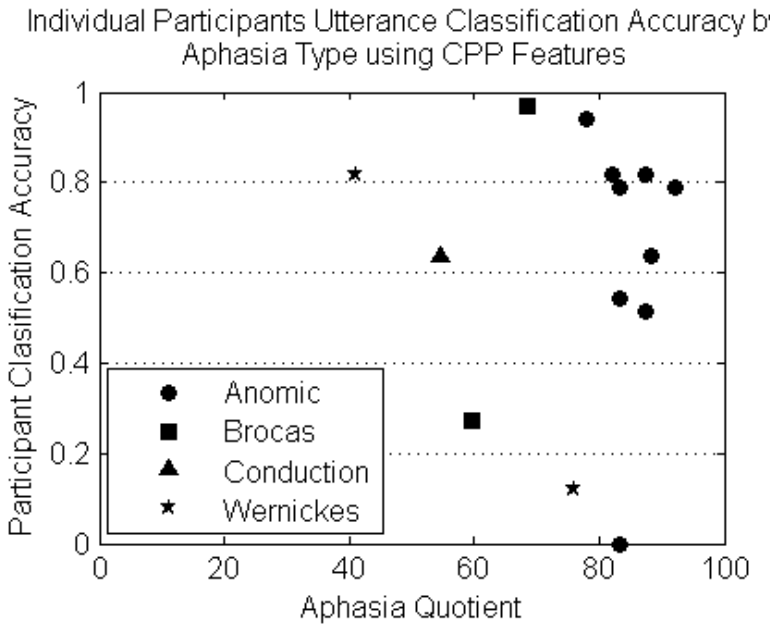

Figure 1: Depression classification accuracy of each participant plotted against their aphasia quotient and aphasia type, © IEEE 2016.

potentially linked to many of the communication and motor disorders experienced by adults with aphasia. Future work will continue this analysis regarding the potential impact on vocal acoustics from motor disorders, especially those measures relating to depression.

In order to determine if the classifier was actually creating a model based on depression or if it was discovering and representing other clinical information besides depression, the classification accuracies of the Cepstral Peak Prominence subgroup were compared to the aphasia type, aphasia quotient, and SADQ-10 numerical score of the participants. Figure 1 shows each participant's classification accuracy against their aphasia quotient and aphasia subtype when assigning the label of depressed/not depressed to each of the utterances for each participant. A score of more than 50% would indicate a majority of their labels were correctly assigned according to the SADQ-10 depression score. Eleven of the 14 participants in this study had the correct label assigned to the majority of their utterances. From Figure 1, it is seen there does not appear to be an indication that the aphasia type or aphasia quotient

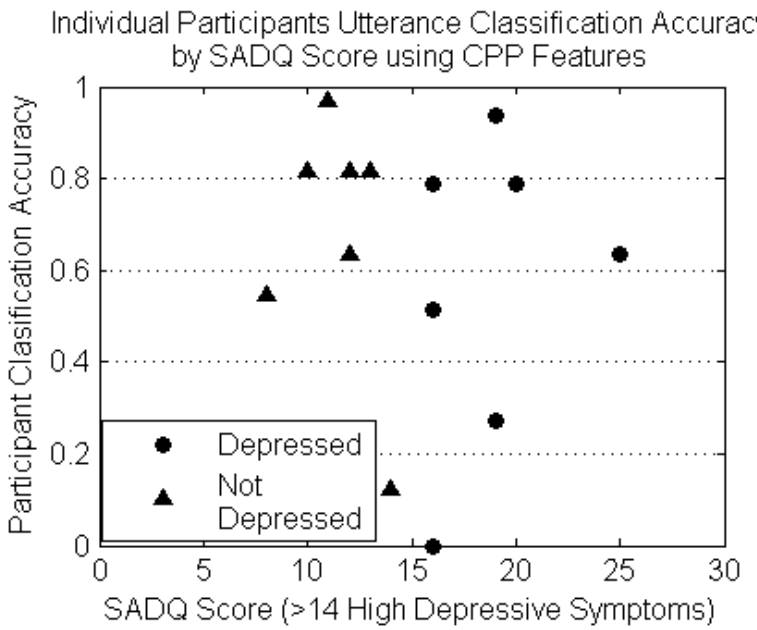

Figure 2: Depression classification accuracy of each participant plotted against their SADQ score and depression label, © IEEE 2016.

impacted the classifier's ability to predict the depression label. Figure 2 shows the classification accuracy in comparison to SADQ-10 scores and depression label. There does not appear to be any indication that the classifier trained on the CPP features classifies either the depressed or not-depressed participants better. An interesting point is that there are three participants whose classification accuracies are below 30%. Two of these three have a SADQ-10 score of immediately below the threshold for no depression (14) or immediately above threshold for depression (15). As mentioned earlier in this study, the SADQ-10 score is a caregiver assessment from which depression may be indicated. However, it is not a precise diagnosis and it is possible that the depression label we are assuming to be a ground truth is false; the results indicate that for the individuals near the depression threshold, their speech has similar feature to those of the opposite class of their depression label.

CHAPTER V

PREDICTING DEPRESSION AND STRESS SCORES IN ADULTS WITH APHASIA

While the work in Chapter 4 presented results of a binary depression classification based on support vector machines (SVM), this work constructs a linear support-vector regression (linear-SVR) model in an effort to predict depression and stress scores from speech. Regression was considered over additional classification methods to create a prediction model that would better emulate how the clinical scores are used by the clinicians as they consider the score itself instead of a binary depression label. The work presented in this chapter was previously published at ICASSP 2017 [114],

5.1 Methodology

From the 26 participants for which speech was collected, 19 participants were selected for analysis by regression of their SADQ-10 scores and 18 were selected for analysis by regression on their PSS (one participant was not able to complete the PSS and was excluded from analysis). A summary of the data used in this analysis is presented in Table 7. Depression was assessed with the SADQ-10 questionnaire. Stress was measured by the Perceived Stress Scale (PSS). The PSS is a fourteen-question assessment to determine the degree to which situations in an individual's life were considered stressful by that individual [74]. For each item, a user can respond with never (item score of 0) to very often (4), for a total of 56 points. The participant PSS scores ranged from 14-40 with higher scores indicating higher levels of perceived stress. The PSS was not designed with severity thresholds or categories, nor can the

Table 7: Summary of demographic and clinical information for the PSS and SADQ-10 regression analysis, © IEEE 2017. *one of the males was excluded from the PSS study since they could not complete the PSS questionnaire.

# Males	12*
# Females	7
Age range	31-70
Aphasia Quotient (AQ)	31.9-99.4
# with AQ >93.8	2
SADQ-10 score range	6-25
PSS score range	14-40

authors find any published study validating proposed categories. The range of values for PSS and SADQ-10 served as the targets for the linear-SVR model.

Prosodic, spectral, TEO and glottal features described in Section 3.2 were extracted from the voiced sections of speech, with low-level descriptors (LLD) statistics calculated at the sentence level as described in openSMILE [82]. A total of 1595 low-level statistics of the features were extracted in MATLAB for each sentence of each participant and normalized using Z-normalization across each individual feature. Feature selection on the full data set as well as each individual feature type grouping was performed first by removing any features with a correlation greater than 0.75, and then using 10-fold cross-validation sequential feature selection to reduce the size of the feature subsets. Only those features that were selected were used to train and test the feature-subset models built using the Support Vector Regression function in MATLAB.

The SADQ-10 and PSS are scored on numeric scales (0-30 and 0-56 respectively) and do not have multiple thresholds representing degrees of severity. Leeds et al. [72] determined a SADQ-10 threshold of 14 as a clinical threshold for the manifestation of depressive symptoms. However, it is difficult to determine how SADQ-10 scores within a range of 1 or 2 points should be interpreted for distinct degrees of depression. In the previous work using SVM classification to detect depression in

patients with aphasia presented in Chapter 4 [109], participants were labeled as depressed or not-depressed based on their SADQ-10 score and the SADQ-10 proposed clinical threshold [72]. Two of the three participants with below 50% accuracy had SADQ-10 scores near the threshold of 14 in the range of 13-16. Cummins et al. [65] recommended excluding any participants that score in the moderate categories of a depression scale due to the ordinal nature of mental state scales. While this exclusion represents an ideal circumstance, often a non-trivial amount of participants will fall in the "moderate" score range, as evidenced by the Aphasia Dataset which contains 7 of 19 participants with +/- 2 of the threshold of 14. As a result, this work will explore the use of a linear support-vector regression model as a predictor of SADQ-10 and PSS values within the current dataset. A linear regression model was examined in [115] on a speech depression dataset without aphasia containing ratings based on the Hamilton Depression Rating Scale (HDRS) [116], a diagnostic which was not designed for persons with aphasia.

To assess the proposed regression tasks, the results and outcomes are reported with respect to mean absolute error (MAE), the R-squared coefficient of determination (R2), and percentage of predicted scores within one standard deviation of the true score (P1SD) [115]. Mean Absolute Error is the average difference between each actual and predicted score, and represents the measure of how close the predicted scores are to the clinical score of interest (SADQ-10 or PSS). The MAE is reported with respect to the sample standard deviation of the diagnostic (SADQ-10 or PSS) in the study by dividing the absolute MAE by the standard deviation of the SADQ-10 (SADQ-σ) and PSS (PSS-σ) scores. R-Squared is used to determine to what extent the variation in the predicted values is determined linearly by the variation in the dependent variable (either SADQ-10 or PSS). P1SD is borrowed from previous work using regression with depression with the HDRS [115] to detect how many predictions were close to the true value.

5.2 Results and Discussion

SADQ-10 scores were predicted for a total of 19 participants (including 2 who had WAB scores that indicated they were non-aphasic). Within the analysis, the SADQ-10 average score was 14.63 and sample standard deviation was 4.97. MAE, R2, and P1SD scores are shown in Table 8 for the various feature types considered in the linear-SVR on the SADQ-10 scores. The MAE results suggest that the average predicted score values were approximately one standard deviation from their true values with HNR and TEO-FM feature subsets providing predictions at slightly less than one standard deviation from the true values. This result is additionally noted in the P1SD scores where HNR and TEO-FM feature subsets achieved 57.5% and 62.1%, respectively. The R2 scores are too close to zero to indicate any significant linear dependency between the predicted values and the true SADQ-10 scores for the HNR and TEO-FM features. Pitch + Jitter and the H1-H2 glottal feature subset exhibited the highest R2 scores of 0.34 and 0.44 respectively.

PSS scores were predicted for a total of 18 participants. Within the population, the PSS mean was 28.55 with a standard deviation of 7.31. Results in Table 9 show TEO-AM performed the best according to the measures considered in the study, with a relative MAE less than one standard deviation of the sample PSS, the second-highest R2 score of 0.29, and 61.1% of sentences predicting the PSS within one standard deviation of the true PSS score. Further analysis of the R2 score of the TEO-AM feature subset results indicated a negative correlation between the predicted PSS scores and the true PSS scores, as shown in Figure 3. These results indicate the linear-SVR model was not sufficient to capture the complexities of the PSS and SADQ scores from the aphasia database.

A correlation analysis was completed to determine any correlations between participants clinical or demographic information and the prediction median, mean, standard deviation, IQR, and/or accuracy. For the top performing PSS feature subset

Table 8: SADQ-10 regression results by feature subtype after feature selection, © IEEE 2017.

Feature Type	MAE(SADQ-10-σ)	R2	P1SD(%)
All	1.24	0.04	46.0%
Pitch + Jitter	1.08	**0.34**	48.4%
RMS-Energy	1.05	0.05	52.3%
LSF + Δ	1.11	0.07	53.0%
MFCC + Δ	1.25	0.04	47.7%
HNR	**0.97**	0.15	**57.5%**
CPP	1.03	0.13	54.4%
TEO-All	1.15	0.03	48.4%
TEO-AM	1.04	0.13	56.8%
TEO-FM	**0.91**	0.00	**62.1%**
TEO-CBarea	1.06	0.03	53.7%
TEO-RMS Energy	1.04	0.00	54.0%
TEO-log Energy	1.04	0.14	50.2%
Glottal-All	1.16	0.12	49.5%
H1-H2	1.16	0.12	49.5%
PSP	1.06	**0.44**	51.9%
HRF	1.07	0.17	51.9%
GLTP	1.21	0.18	44.9%

Table 9: PSS regression results by feature subtype after feature selection, © IEEE 2017.

Feature Type	MAE(PSS-σ)	R2	P1SD(%)
All	1.51	0.12	36.8%
Pitch + Jitter	1.05	0.18	54.4%
RMS-Energy	**0.94**	0.12	**61.4%**
LSF + Δ	1.17	0.11	43.9%
MFCC + Δ	1.33	0.11	43.9%
HNR	1.02	0.15	55.8%
CPP	0.97	0.07	54.7%
TEO-All	1.01	0.00	55.3%
TEO-AM	**0.94**	**0.29**	**61.1%**
TEO-FM	1.06	0.25	53.2%
TEO-CBarea	0.97	0.00	55.0%
TEO-RMS Energy	1.01	**0.40**	56.1%
TEO-log Energy	0.98	0.09	59.4%
Glottal-All	1.05	0.02	52.9%
H1-H2	1.05	0.02	52.9%
PSP	1.01	0.24	52.9%
HRF	**0.89**	0.00	59.9%
GLTP	1.00	0.01	53.5%

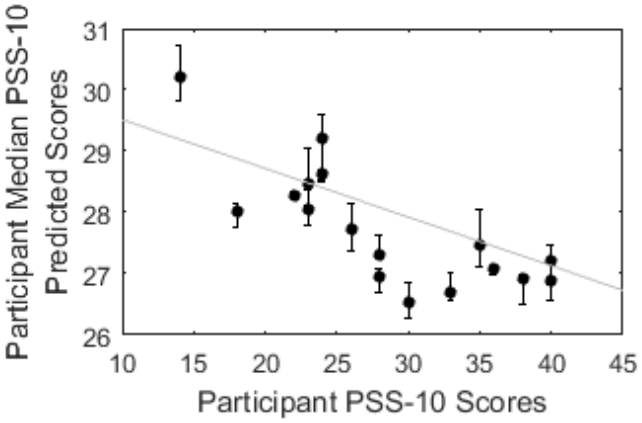

Figure 3: 1st, 2nd, and 3rd quartiles and linear regression line for prediction of PSS scores for TEO-AM feature group © IEEE 2017.

(TEO-AM), no statistically significant correlation was found between the calculated measures and any of the clinical or demographic information. Similarly, no other feature subset results correlated significantly between the predictions and any clinical or demographic information available for analysis. The lack of significant correlations indicates that the models do not appear to perform significantly better or worse due to any specific trait of the participants. Instead, the linear-SVR performance is likely hindered by the complexities of the diverse population from which the dataset was collected.

CHAPTER VI

ANALYSIS OF CHANGE IN AFFECT IN ADULTS WITH APHASIA

The participants available for analysis from the Aphasia Dataset vary greatly with respect to age and gender. Many participants with aphasia are also diagnosed with dysarthria or apraxia, motor disorders impacting the speech produced. As an additional complication, the PSS and SADQ-10 scores are based off of the participants emotional state throughout the past month—it is often possible for a participant who is depressed or stressed to appear happy or calm during the interview time period which could change and impact the emotional content of their speech. Furthermore, the current clinical labels for depression and stress are based off of the SADQ-10 (a survey completed by the caregiver) and the PSS (a survey not modified for the aphasic population, currently read out loud to the participant with explanations or substitutions of words provided as necessary).

The results of Chapters 4 and 5 have shown that the detection and/or prediction of depression based on the SADQ-10 score or perceived stress based on the PSS score is challenging. It was hypothesized that the short-term affective state could be dominating the vocal acoustics extracted. The acute-nature of affect and emotional states would make building a predictive model for affect challenging due to the time-changing nature of affect. As an alternative approach, the work presented in this chapter attempted to correlate changes in vocal acoustics to reported changes in affect collected from pre- and post- recording assessments of affect. To accomplish this, two tasks were performed:

1. Analyzing the correlation between collected SAM, Stress-Scale, and WAB scores,

providing information regarding the consistency of the self-reported scores and the potential implications of aphasia on affect, and

2. Determining the acoustic measures which correlate to changes in affect, testing the hypothesis that short-term affective states are more prevalent in the vocal acoustics than long-term stress or depression in adults with post-stroke aphasia.

The work presented in the remainder of this chapter has been submitted for publication to the Journal of Speech, Language, and Hearing Research [117].

Participants from the Aphasia Database were included on the condition of proper audio recordings and confirmed history of stroke. This resulted in a total of 20 participants (included the two who with scored as higher than the threshold for aphasia diagnosis, presenting in the non-aphasic range on the WAB) for the affect analysis. The two participants who scored above the WAB cut-off were included in the analyses as they self-identified as having aphasia and were referred to the study due to their aphasia. Speech was limited to spontaneous-speech of the phrase- or sentence-length only. The full feature set as detailed in Section 3.2 (prosodic, spectral, glottal, and TEO) was extracted from each sample.

6.1 Correlation Analysis

Due to the variety of labels collected from both caregiver assessments as well as self-reports from the participant, the correlation between similar affective scores was necessary to validate the consistency in responses between the various affective assessments. Laures-Gore et al. recently published the results comparing the Aphasia Depression Rating Scale (ADRS), SADQ-10, and PSS scores for 25 of the participants included in the original data collection [73]. Laures-Gore et al. did not find any significant correlations between the ADRS or SADQ-10 scores and the other demographic information considered in their analysis (age, WAB-AQ, or time post-onset from stroke). In this work, further correlation analysis using Pearson product-moment

correlation was completed with a larger subset of the available clinical scores for the 20 participants analyzed for affect.

Selected results of the extended correlation analysis are shown in Table 10. High correlation values between the SAM-arousal and Stress-Scale scores were found between both pre-recording scores and both post-recording scores (pre: r=0.712, post: r=0.873), verifying that participants were consistent with their selection of pictures relating to arousal and stress regardless of a 1-5 or 1-7 scale. Moderate correlations were found between SAM-valance and SAM-arousal scores, indicating participants may not have been able to isolate these aspects of their emotion as they related to the pictures shown before them. There was a moderate to strong correlation between each subtest of the WAB and the overall aphasia quotient, an expected finding as the WAB Aphasia Quotient is based off a linear scoring of the individual subtests. However, it is interesting to note that the spontaneous speech sub-score and the word finding sub-score were the most highly correlated with the Aphasia Quotient (AQ). The spontaneous speech score is weighted more heavily than other sections when computing the Aphasia Quotient and would influence the AQ more, but the word-finding score is not.

6.2 Methodology

During this study, individuals may have experienced a change in their affective state depending upon their personal response to the testing procedure, which could cause a change in their speech acoustics. Participants who maintained the same affective scores from the pre- and post- labels should have minimal differences in the statistics computed on their speech throughout the entire recording process, while participants who reported a change in affective scores should have differences between the statistics computed from the beginning and end of the recording session. Pre- and Post- scores from the SAMs and the Stress-Scale provide a measurement from which to calculate

Table 10: Correlation matrix for clinical and affective scores of the Aphasia Dataset

	SAM Arousal-Post	Δ SAM Arousal	SAM Valance-Pre	SAM Valance-Post	Δ SAM Valance	Stress Scale-Pre	Stress Scale-Post	Δ Stress Scale	PSS	WAB Sub-Spon	WAB Sub-Word Finding	WAB-Aphasia Quotient
SAM Arousal-Pre	0.35	-0.45	**0.82**	0.33	-0.37	**0.71**	0.20	-0.42	0.52	0.24	0.16	0.22
SAM Arousal-Post		0.67	0.29	0.54	0.30	0.35	**0.87**	0.47	0.31	-0.31	-0.23	-0.33
Δ SAM Arousal			-0.38	0.25	0.58	-0.23	0.67	**0.79**	-0.11	-0.49	-0.42	-0.49
SAM Valance-Pre				0.45	-0.40	**0.70**	0.19	-0.42	0.40	0.20	0.31	0.22
SAM Valance-Post					0.64	0.42	0.54	0.12	0.52	-0.08	0.01	-0.18
Δ SAM Valance						-0.18	0.39	0.50	0.17	-0.25	-0.25	-0.37
Stress Scale-Pre							0.33	-0.37	0.56	0.05	0.23	0.11
Stress Scale-Post								0.61	0.39	-0.36	-0.18	-0.34
Δ Stress Scale									-0.13	-0.36	-0.35	-0.39
PSS										-0.08	-0.04	-0.13
WAB Sub-Spon											**0.88**	**0.93**
WAB Sub-Word Finding												**0.93**

a change in affective state. The reported change in affective state can be correlated to changes in statistical distributions of the features to determine to what extent the speech acoustics are representative of affective state instead of long-term stress or depression.

The first and last ten sentences from the 20 participants who had complete recordings and had record of confirmed stroke were selected for analysis in Matlab. Limiting the number of samples considered at the beginning and end of the interview to ten ensured minimal time elapsed between the SAM and Stress-Scale assessments and when the sentences were recorded, minimizing the likelihood that the self-reported affective labels would have changed. All 1595 statistics (including the aforementioned prosodic, spectral, glottal, and TEO features) were extracted for each sentence. Looking at all 1595 features at once would cause the dimensionality to be too large to effectively analyze, while it was unlikely that any single feature would be able to perfectly represent the change in affective scores for all 20 people. However, a small set of features, when looked at together, would be an appropriate subset from which to understand the reported changes in affect.

6.2.1 Feature Selection Process

Figure 4 shows a flow-chart of the small-group feature selection process, which included grouping participants by their delta-SAM scores, calculating the Wilcoxon Rank Sum probability for each feature's distribution, calculating feature scores that indicate how well a feature represents the changes in affect, and then the feature selection processes based on the feature scores.

Participants were grouped according to the type of change self-reported in the individuals affective scores: SAM valance, SAM arousal, and the Stress-Scale assessments. The change in affective score was calculated by subtracting the pre-recording score from the post-recording score. Based on the sign of the change in affect score,

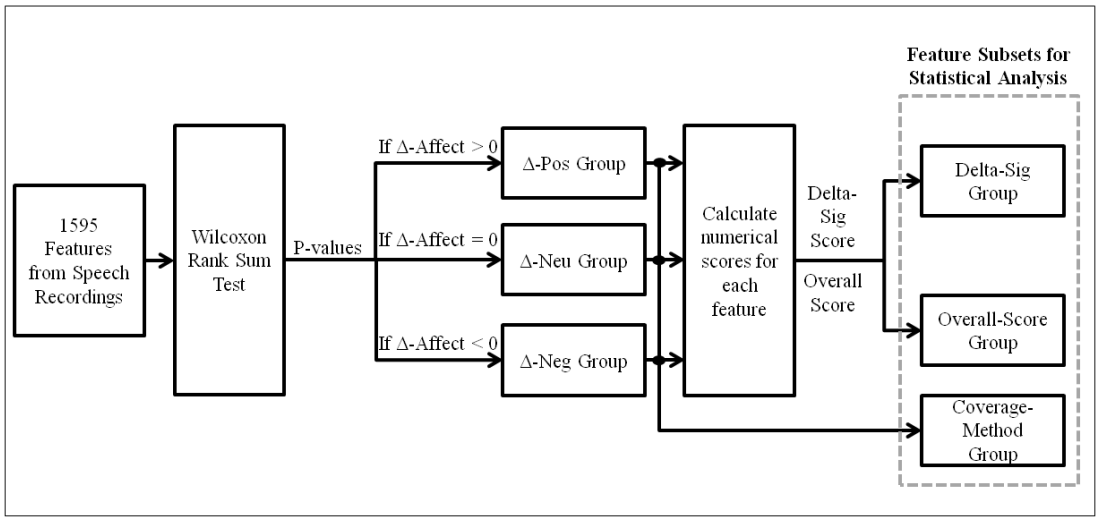

Figure 4: Flowchart visualizing the feature-selection process from 1595 objective speech acoustics into three types of sub-groups, ranging in size from 2-54 features each

participants were assigned to one of three groups for each affective label, which will be referred to as Delta-Negative (Δ-Neg), Delta-Neutral (Δ-Neu), and Delta-Positive (Δ-Pos). The Δ-Neg group had a lower self-assessment score at the end of the recording procedure than their initial score, the Δ-Neu group had the same score as their initial score, and the Δ-Pos group had a higher self-assessment score at the end than their initial score.

A change in affect should manifest itself as a distinct change in the distribution of the vocal acoustics. The two-sided Wilcoxon Rank Sum (WRS) nonparametric test uses the null hypothesis that the samples from two given distributions came from a single distribution. The WRS test was completed for each participant and each feature to determine, at the participant level, which features likely came from the same distribution and which features rejected the null hypothesis based on the p-value. This procedure has been used in prior vocal acoustic work to assess the extent of voice impairment by comparing the calculated features of speech from adults with Parkinson's Disease to speech from healthy controls [91].

Using these groupings and the calculated WRS p-values, two numerical scores were

calculated for each feature. The Delta-Sig Score was the total number of participants in both the Δ-Pos and Δ-Neg groups who had a significant p value ($p<0.05$) that rejected the null hypothesis of the WRS test. A higher Delta-Sig Score indicates more rejected null hypotheses for individuals across the same feature when there was an expected difference in distributions of the features as indicated by the reported change in affect. The Overall Score was calculated by subtracting the number of participants in the Δ-Neu group who had a significant p-value from the Delta-Sig Score. This acted as a negative weighting for features receiving significant p-values in the Δ-Neu group, participants from which a rejection of the null hypothesis would not be explained by a change in affect. Three feature selection methodologies were completed using these scores.

The first feature selection, referred to as the Delta-Sig Group method, ranked features first according to the Delta-Sig Score (highest first) and then according to the Overall Score (highest first). Any feature that had an Overall Score of 0 or below was removed regardless of the Delta-Sig Score. Three groups of features of increasing size (DSG1, DSG2, DSG3) were then created by selecting the top features based on natural cutoffs in the data discerning between different Delta-Sig scores. Groups of the DSG method had sizes ranging from 4 to 44 features. The DSG method can be summarized as selecting the features that were most likely to come from two distinct distributions in those participants who reported a change in affect.

The second type of feature selection, referred to as the Overall-Score Group method, ranked features first based on the Overall Score, and secondarily on the Delta-Sig Score. As before, three groups of features of increasing size (OSG1, OSG2, OSG3) were selected based off of natural cutoffs in the data that allowed for various group sizes to be considered. Groups of the OSG method ranged from 5 to 54 features. The OSG method attempted to select the features that were most likely to come from two distinct groups only for those participants who reported a change in

affect, penalizing features that had a small p-value for participants with no reported change.

The last type of feature selection, referred to as the Coverage Method, attempted to optimize the features selected to ensure that there was at least one feature per person in the Δ-Neg and Δ-Pos groups for whom a significant p-value was present, while minimizing both the number of features selected and the number of participants in the Δ-Neg group with a significant p-value. The feature subsets will be referenced as CMG1, CMG2, and CMG3, where the number indicated the minimum number of features with small p-value for each participant selected. The CMG method resulted in sets of features with group sizes ranging from two to six features. While the manual optimization was potentially not the optimal solution, it still created smaller group sizes than either the OSG or DSG methods. The CMG method focused on selecting the minimal-number of features while still ensuring every participant with a reported change in affective score was represented by the features selected with significant WRS p-values.

6.2.2 Correlation of Distribution Distances to Change in Affect

In total, 27 feature subsets were created using the aforementioned statistical methods (3 methods of 3 different sizes for the 3 affective assessments). To determine to what extent the feature subsets explained the changes in the self-assessment scores for the various emotions within a single individual, various measures were calculated between the First-10 and Last-10 samples of each individual in the N-dimensional subspace, where N is the number of features in each subset. Both the Euclidean distance and the cosine similarity were calculated between each point of the First-10 samples and each point of the Last-10 samples, resulting in a total of 100 distance calculations. Then, the statistics of minimum, maximum, and mean values were found for both the Euclidean distance and the cosine similarity score. Additionally,

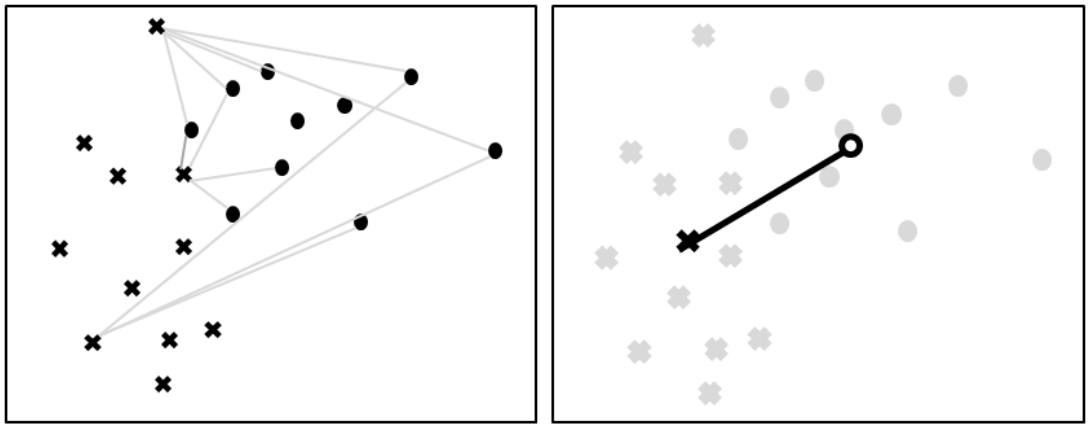

Figure 5: Visualizations of the distance metric calculations between the first-10 and last-10 samples of each participant. The two classes are identified as x and o respectively, representing the first-10 and last-10 recordings of each person. Left image shows some of the distance measures calculated between points of the different classes, from which the average, minimum, and maximum distances would be calculated. Right image shows the calculated average of each class, from which the between-class-mean distance was calculated

the average samples of both the First-10 and Last-10 groups were found by averaging all 10 samples. From these two mid-points, the class average Euclidean distance and cosine similarity were calculated. In total, four score-statistics (minimum, maximum, mean, and between-class-average) were analyzed for both the Euclidean distance and cosine similarity scores. Figure 5 shows an example of the statistics calculated from two sample distributions for visualization purposes.

To compare the calculated statistics to the change in affect, each participant's calculated distance measure statistics were correlated to three different measures associated with the affective score. The three measures were the numerical change in affective score, the sign of the change in affective score (represented as +1 if positive, -1 if negative, or 0 if no change), and the presence of a non-zero change in affective score (1 if participants reported any type of different affective score and 0 if participants reported the same affective pre- and post- scores).

6.3 Results and Discussion

Figure 6 shows the calculated feature values for the first and last 10 samples of speech across all 20 participants with the features selected by the Coverage Method for the SAM-Valance label. This feature subset was selected for visualization due to its simplicity with just two features, represented in two-dimensions. While the specific features chosen (kurtosis of the delta of the 7-MFCC value and the linear quadratic error of the 7th LSF coefficient) do not have any meaning that can be translated to a directly identifiable trait or characteristic of the speech, they were selected based off of their significant p-values as determined by WRS procedure that would potentially identify changes in affective scores. The plots boxed with solid black boxes, light grey boxes, and no boxes identify participants in the Δ-Neg group, the Δ-Pos group, and the Δ-Neu group respectively. Visually, it can be seen that certain participants (e.g. 16 and 21) had a clear change in their sample distribution between the first and last 10 samples. However, not every participant with a change in SAM-Valance score (as indicated by either a gray or black box around the plot) had visible speech acoustic changes that confirmed their self-reported change in affect. Additionally, some participants appear to have changes in speech acoustics, but did not report a change in affect. The speech acoustics extracted in this work do not correlate perfectly to the self-reported affect scores, and suggest that changes in affect manifest themselves as changes in speech acoustics differently in each individual's speech acoustics. As such, the moderate correlations reported here are of significant interest, with the results showing promise in being able to detect changes in affective states in individuals with aphasia.

Pearson Correlation coefficients were considered significant with a two-tailed p-value of 0.05 if $|r| > 0.44$ for N=20. Additionally, this criteria satisfies the definition of moderate correlation set by Evans of $|r| > 0.4$ [118]. The Euclidean distance and

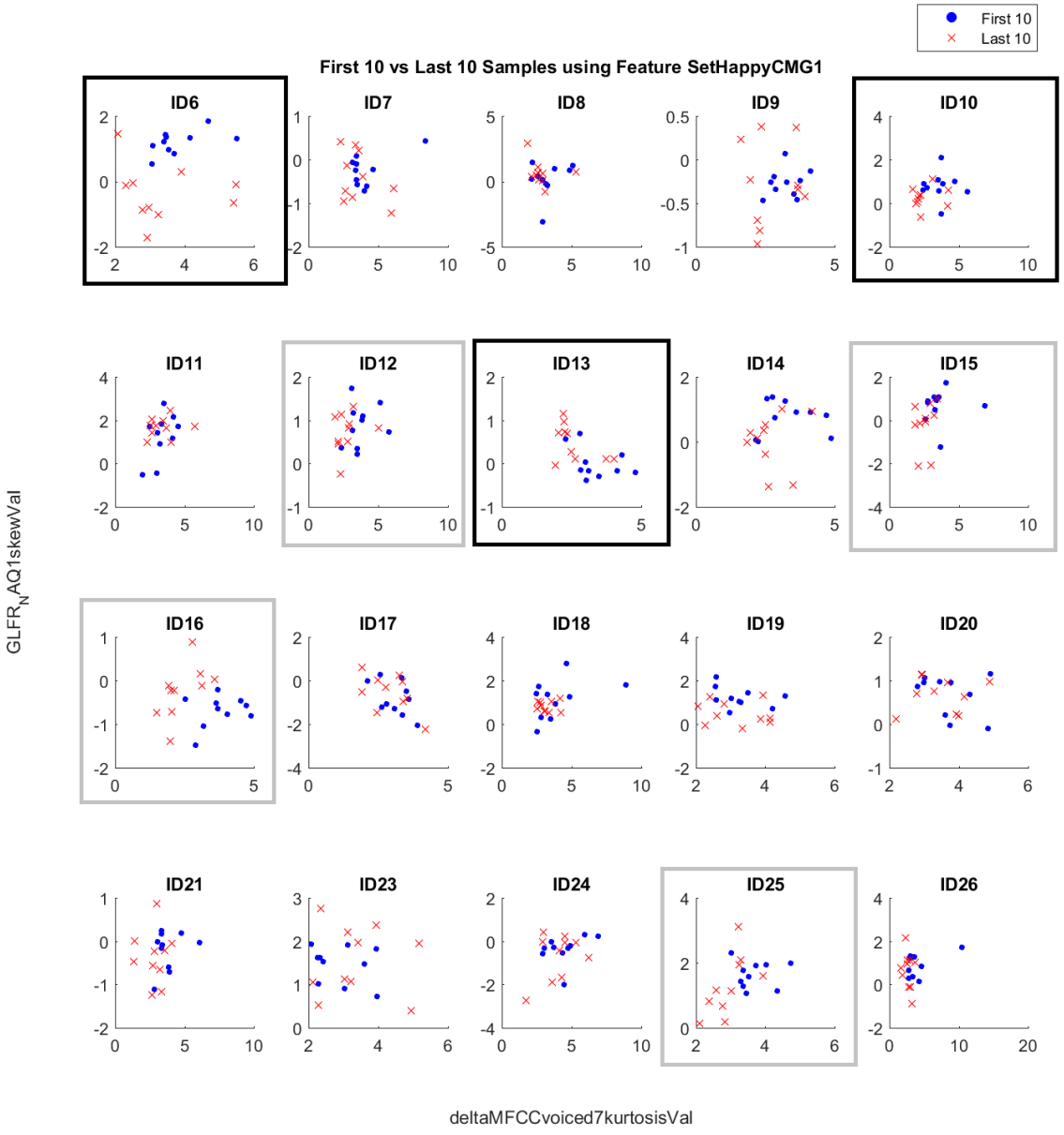

Figure 6: Distributions of Coverage Method feature values for the first ten and last ten sentences of each person for the SAM-Valance test scores. Plots with a solid black outline displayed a negative change in SAM-Valance scores, plots with a solid gray outline displayed a positive change in SAM-Valance scores, and plots with no outline did not report a change in SAM-Valance scores.

Cosine similarity had the most number of significant correlation coefficients, irrespective of the affective assessment, when correlated to the binary label representing any change in affective score. This was the broadest category considered with the most balance in the number of participants that fell into the two groups of interest (change in emotion, +1, or no change in emotion, 0). Attempting to correlate to the sign of the change (3 possible values) or the numeric value of the change itself (up to 13 possible values if Stress Scale, or 9 possible values if SAM scores) provides less generalization and worse performance due to the limited number of participants available for analysis since fewer participants would be spread out across a larger range of values.

Of the 9 feature subsets considered for analysis for each of the 8 distance measures and all 3 affective assessments, a total of 15/216 or 6.9% had significant correlations to the presence of change in emotion, as compared to 5.6% for correlations to the changes in the direction of the of the delta change (positive, negative, or neutral) and 1.9% for correlations to the numeric delta-scores. Thirteen of the fifteen significant correlations to presence of change were measures correlated to the Stress-Scale instead of valance or arousal SAM scores, potentially indicating either that changes in stress manifest themselves more clearly in the speech of patients with aphasia than possible changes in arousal or valance, or that it was easier for participants to identify their emotion on the Stress-Scale than the SAM. Table 11 shows, for each affective assessment, the best feature subset and the corresponding correlation coefficients between the distance measures and the presence of affective change. The average distances or the between-class average distance tended to have higher correlation coefficients across both the Euclidean and Cosine distance measures. This is likely because averages better handle potential outliers compared to the minimum or maximum. The presented statistics show that carefully select features groups reveal statistically significant correlations to changes in affective state.

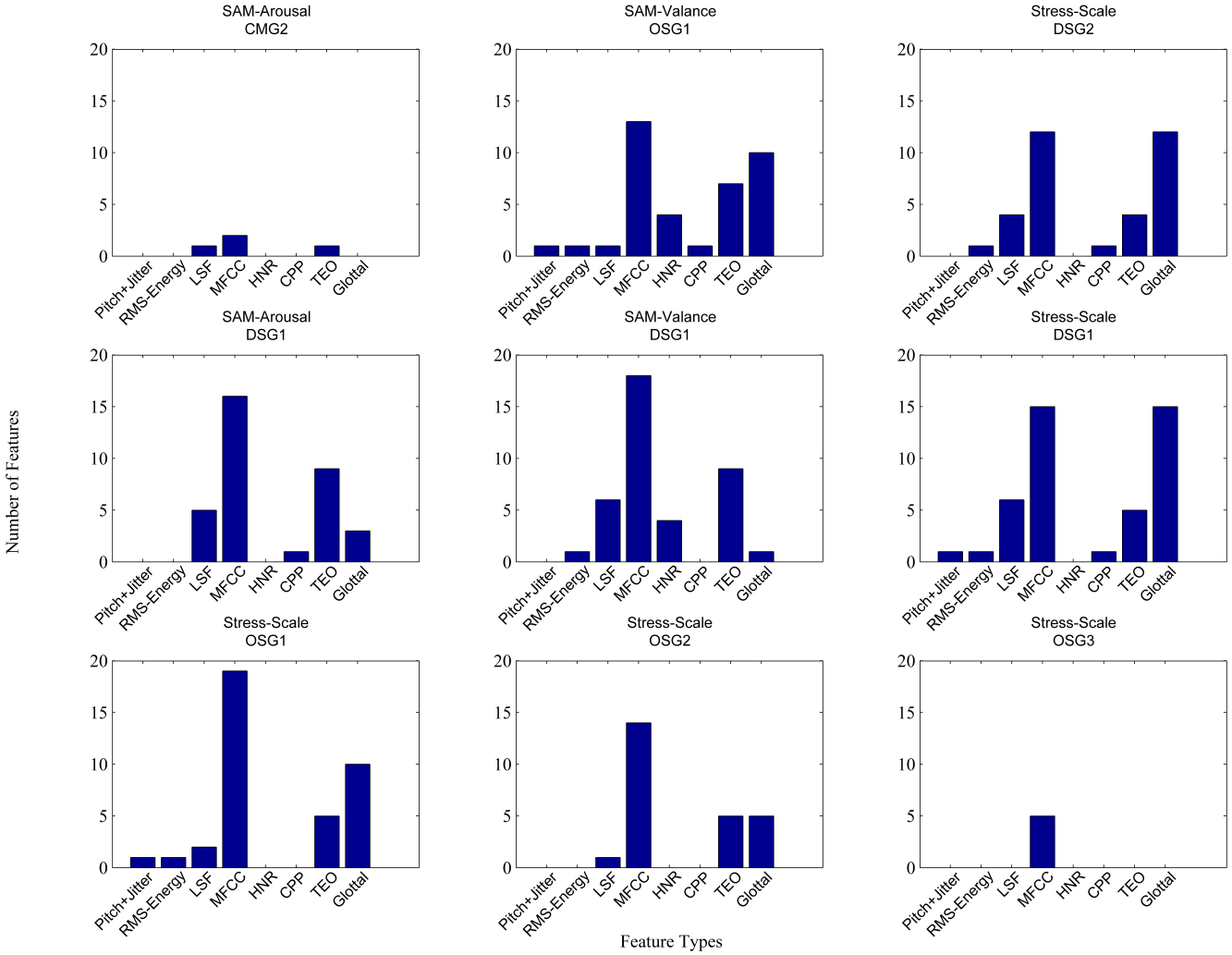

Figure 7: Distributions of the number of each feature type selected for various feature subsets. First row are the features selected in the subsets presented in Table 11. Second row are the feature subsets for DSG1 method across all three affective assessments. Third row is the OSG method for Stress-Scale with all three different sizes.

Table 11: Correlation coefficients of the feature subsets with the largest number of significant correlations with the distance measures for each affective label, where significant is specified as $|r|>0.44$, for 2-sided probability of $p<0.05$, N=20

Affect Label	Feat. set ID	Num. of Feats.	Min Eucl. Dist	Max Eucl. Dist	Avg Eucl. Dist	Eucl. Dist. to mean	Min Cos. Dist	Max Cos. Dist	Avg Cos. Dist	Cos. Dist. to mean
Arousal	CMG2	4	0.28	0.10	0.220	-0.06	0.34	0.17	0.41	0.43
Valance	OSG1	39	0.19	0.29	**0.54**	0.21	-0.03	-0.39	-0.36	-0.20
Stress Scale	DSG2	34	-0.06	**0.53**	**0.45**	**0.51**	**0.46**	**0.49**	**0.53**	**0.49**

There was no obvious trend across the feature selection methods (OSG, DSG, or CMG) to suggest better results when selecting more or fewer features. However, all three methods already drastically reduced the number of features considered from the original 1595 low-level descriptors down to feature sets with sizes ranging from 2 to 54. Additionally, there was no clear distinction in a specific feature selection method resulting in better results when considering the different affective assessments or the affective-change categories of the correlation methods. Figure 7 shows selected histograms visualizing the number of each feature type selected by the feature selection algorithms for some of the feature groupings. It can be seen that the MFCC features were selected in 100% of the feature-selection tasks and were also selected most often, with over 50% of the total features selected belonging to the MFCC type. Glottal features were the next highest, with 62% of feature groups selecting at least one glottal as a total of 19.8% of all features selected. LSF and TEO features were selected in about half of the subsets. Pitch features were only selected in 14% of the 27 subgroups. While often indicative of emotion, it is possible the pitch already fluctuated from sentence to sentence to such an extent that the potential affective changes did not influence the distances calculated from the pitch statistics.

While a direct comparison of the features' abilities to predict changes in affect

compared to the features' abilities to predict changes in stress or depression is not possible due to the difference in the methodology performed (using the data to predict a single clinical label versus correlating the differences in the data to the difference of two acute temporal labels), the results suggest that changes in affect often influence the features used in the speech analysis of this study. The prominence of acute affective state in the recordings collected for this work could explain the disparities between the higher accuracies reported in the literature for affective processing in non-aphasic speech and the results reported in the prior work detecting stress and depression in adults with aphasia. Since the speech of adults with aphasia is often described as effortful and the recording process was an extended session, it is possible that the participants in this study were more likely to experience changes in acute affect as the study progressed compared to the participants in a depression study on persons without aphasia. However, no formal comparison or conclusions can be made regarding this claim as acute affective states have not been reported for speech analysis studies of clinical depression in adults without aphasia.

6.4 Aphasia Database Limitations

The presented work on affect analysis was not easy due to multiple challenges, many of which can be attributed to the data collection. The small database was limited by eligible participants and their caregivers who are willing and able to successfully complete the entire data-collection process and clinical diagnosis procedure, as well as the extensive time-commitment to collect the data. As such, there was a large range in clinical and demographic information for those participants who chose to participate, making generalizations of models complicated and challenging. Additionally, the data may be considered limited to a single snap-shot of participants mood during a single recording session, while speech analysis for detection of long-term clinical depression and stress often rely on databases that have multiple recordings over multiple months

for each participant. As an analogy, this would be similar to clinicians being asked to use SAM scores to detect long-term stress and depression instead of the SADQ-10 and PSS assessments; when detecting depression, caregivers and the individuals are usually asked to comment on behaviors over a time-period of up to a month and not the emotional state from just a single day. Perceived stress (as measured by the PSS) and depression (as measured by the SADQ-10) are not acute conditions, and as such, acute assessments should not be the primary diagnostic tool. Future studies of affect in adults with aphasia, especially stress and depression, should strongly consider a multi-session recording and assessment process to allow the speech collection to extend over a longer time-period.

CHAPTER VII

AUTOMATIC ANALYSIS OF DYSARTHRIA PRESENCE USING CROSS-DATABASE MODELS

The work in Chapters 4, 5, and 6 analyzed the Aphasia Database for depression classification, prediction of SADQ-10 and PSS scores for depression and perceived stress, and feature distributions correlations to changes in acute affect. While most of the participants in the Aphasia Database presented with dysarthria and apraxia of speech, the impact of these motor disorders was not considered in the prior due to the limited number of participants available. Attempting to learn a model which would compensate for dysarthria as well as analyze stress, depression, or affect would not be possible using the Aphasia Database due to the varying clinical and demographic labels of the participants. An alternate approach would be to build a model using other data representing dysarthria, and then apply it to the Aphasia Database. A cross-database method is investigated to overcome the small database sizes present in clinical data. A subset of the work presented in this chapter was submitted for publication and is under review at the time of publication of this dissertation [119].

The goal of the work in this chapter is to begin the analysis necessary to design a clinical tool for the automatic detection of dysarthria from vocal acoustics of continuous speech. An automatic dysarthria diagnostic system could be used in a clinical setting on its own, or as a pre-processing tool for research related to affect in language disorders, particularly as it relates to the analysis of aphasic speech and affect presented in earlier chapters. This is an important initiative since dysarthria is a co-occurring motor disorder with up to 50% of adults with aphasia also diagnosed as having dysarthria [14]. However, most of the speech databases including

participants with dysarthria have a small number of participants. In an attempt to overcome these small sizes, this work utilizes cross-database training for the detection of dysarthria. There is little prior work which uses multiple corpora for classifying dysarthric speech from non-dysarthric speech or identifying dysarthria subtype and severity; one example is work by Orozco-Arroyave et al., which focuses on detecting hypokinetic dysarthria for Parkinson's disease across different languages [90].

Detection of dysarthria from vocal acoustics has occasionally been studied in recent years. Classification results from analysis of vowels by Mekysaka et al. achieved accuracy results ranging from 72-92% depending on the vowel and spoken conditions [120]. DeCicco and Patel used machine learning in vowels from children with dysarthria to automatically detect pitch and duration manipulations for use in alternative and augmentative communication [121]. Vyas et al. extracted MFCCs, skewness, and formants from 40 speech utterances evenly split for training and testing of dysarthria vs controls from UA-Speech [122]. An accuracy of .98 for all speech and .87 for dysarthria speech was reported. Intelligibility classification based on the UA-Speech dataset has been considered by Martinez et al. achieving results of 0.60 for weighted precision and 0.44 for weighted recall in a leave-one-subject-out, 4-class setting [44]. Paja and Falk utilized a two-stage analysis for spastic dysarthria detection and reported results of .83-.95 on the UA-Speech dataset using prosodic, linear-predictive, MFCC, HNR, and Glottal-to- Noise Excitation (GNE) features [25].

The UA-Speech dataset [35] is one of the largest, freely-available datasets of dysarthric speech and is used to establish a comparison of the presented work to prior work. The inclusion of speech from healthy controls allows a model to be built to detect the presence of dysarthria. Previous work on the UA-speech database has been able to take advantage of the singular type of dysarthria presented in the data as well as the uniformity in the type of speech collected. However, the UA-Speech dataset is not sufficiently diverse to create a clinical dysarthria detection tool: the

majority of speakers have spastic dysarthria originating from cerebral palsy. To expand the applicability of the proposed tool, the models trained on UA-Speech are tested on the Atlanta Motor Speech Disorder Corpus (AMSDC) [79], containing a wider variety of dysarthria types and etiology. The cross-database training for this study uses the UA-Speech dataset as the only training set. Without speech from healthy controls, the AMSDC is not suitable to be used to train the detection model as there would be no counter-examples to the dysarthric speech.

7.1 Methodology

From the AMSDC, a subset of 57 participants who presented with dysarthria was selected for analysis in this work. The AMSDC contains continuous speech samples that needed to be manually segmented for this pilot study in making more direct comparisons with UA-Speech. An estimated 30 words were manually segmented from the AMSDC sentences with contrasting speech emphasis to match the length of the recordings in the UA-Speech dataset. These words were identified to include different parts of speech and avoided the stressed syllable of the phrase, but were not able to be selected to be phonemically balanced. The manual segmentation of the speech from the contrasting stress emphasis portion of the AMSDC resulted in some participants having a maximum of 12 words segmented due to low intelligibility. In order to create an even comparison for cross-database training, all participants were restricted to 12 words for training/testing. Therefore, a reduced-UA-Speech dataset was used with 12 words per person (taken from the same microphone of the original 8-microphone array for each individual) or 336 total words for training purposes.

Voiced sections of speech were identified from the word-length recordings of both datasets. The prosodic, spectral, TEO and glottal features described in Section 3.2 were extracted, and statistics based on those reported in openSmile (i.e., average, min, max, slope) were calculated at the word or sentence level [82]. A total of 1595

low-level statistics of the features were extracted in MATLAB for each response and normalized across each individual feature by subtracting the global mean and dividing by the global standard deviation. Initial feature reduction involved removing features with a correlation greater than 0.75. Gaussian noise on the order of 10^{-6} was added to the data to ensure the matrix was well conditioned for the machine learning. A 10-fold cross-validation sequential forward feature selection (SFFS) was used to further reduce the size of the feature subsets. In general, the feature selection strategy selected between 3-6 features for each experiment. It is possible by changing the criteria function of the SFFS algorithm, more features would have been included and results could have been improved. However, parameter searches to tune the feature selection algorithm or SVM parameters were not of interest in this work since the focus was on the impact of cross-dataset setup on classification results. Only the selected features were used to train and test the feature-subset models built using the Support Vector Machine function with a radial-basis function (RBF) kernel in MATLAB.

It is common to have small data-sets when working with clinical data, especially those of speech language or motor disorders. To compensate for the small numbers of participants, multiple databases can be used to train and test the algorithms. Alghowinem at al. and Tahon et al. are two examples that have used multiple datasets when detecting depression or classifying emotions in populations not affected by language or motor disorders [123, 124]. Both studies observed performance declines when a model was trained on one dataset and tested on another, likely due to differences in the recording environments. To accurately compare the training and testing of multiple datasets in this work, a baseline is established by training and testing a model on only the UA-Speech dataset, and then applying and testing the model on the AMSDC. A leave-one-subject-out cross validation strategy was conducted per subject to determine the classification accuracy for the detection of dysarthria presence from the controls in the UA-Speech dataset. Selected classification results from using

the reduced-UA-Speech dataset are presented in the first column of Table 12 at the word and participant level. The word level accuracies are calculated by comparing each word's predicted label of dysarthria or no-dysarthria to the known condition of the individual. The participant-level accuracies are the percentage of participants for which at least 50% of the individual's words were classified correctly.

The main goal of this work was to perform cross-database training and testing as an initial step to create a dysarthria classification tool. As mentioned previously, the act of training a model on the AMSDC and testing on the reduced-UA-Speech would not be appropriate in this work due to a lack of healthy-control speech included in the same recording environment as a part of the AMSDC. Instead, a model for the detection of the presence of dysarthria was trained on the reduced-UA-Speech dataset, and then applied to the AMSDC. It could be predicted that testing the model on speech most similar to what the model was trained with should produce higher accuracies. As such, the first iteration compared the accuracy of detecting dysarthria on those participants of the AMSDC presenting with spastic dysarthria, the type of dysarthria most commonly represented in the UA-Speech dataset. Word and participant-level classification accuracies for each feature-set are presented in the third and fourth columns of Table 13. Next, the accuracies were calculated amongst the participants who had either spastic or flaccid dysarthria, the two types of dysarthria represented in the UA-Speech dataset. These results are presented in the fifth and sixth columns of Table 13. The last experiment applied the dysarthria detection model to all of the AMSDC, regardless of dysarthria type. These results are presented in the two right-most columns of Table 13.

Table 12: Classification accuracy of predicting dysarthria in the UA-Speech Dataset based on a model trained on the UA-Speech dataset.

Feature Type	Average Number of Features Selected	Word-level Classification Accuracy	Word-level Precision	Word-level Recall	Participant Classification Accuracy
All	5.79	0.661	0.706	0.706	0.750
Pitch + Jitter	3.79	0.670	0.725	0.689	0.786
RMS-Energy	3.64	0.648	0.693	0.639	0.750
HNR	3.04	0.673	0.700	0.706	0.786
All-Prosodics	4.82	0.753	0.824	0.728	0.923
LSF + Δ	5.32	0.568	0.583	0.561	0.607
MFCC + Δ	4.96	0.670	0.709	0.650	0.786
CPP	3.75	0.714	0.736	0.728	0.893
All-Spectral	5.29	0.607	0.648	0.633	0.679
TEO-AM	3.93	0.664	0.722	0.694	0.786
TEO-FM	3.46	0.592	0.624	0.656	0.679
TEO-CBarea	4.82	0.711	0.763	0.733	0.857
TEO-RMS Energy	3.39	0.693	0.696	0.739	0.786
TEO-Log Energy	4.46	0.714	0.715	0.739	0.857
All-TEO	5.89	0.726	0.764	0.739	0.893
H1-H2	3.32	0.705	0.723	0.694	0.821
PSP	4.28	0.622	0.623	0.689	0.714
HRF	3.89	0.610	0.613	0.650	0.607
GLTP	5.54	0.705	0.766	0.744	0.821
All-Glottal	5.57	0.711	0.754	0.750	0.857

Table 13: Classification accuracy of predicting dysarthria in the AMSDC from a model cross-trained on the reduced-UA-Speech dataset.

Feature Type	Number of Features Selected	Spastic-only AMSDC		Spastic and Flaccid AMSDC		All AMSDC	
		Word-level Accuracy	Participant-level Accuracy	Word-level Accuracy	Participant-level Accuracy	Word-level Accuracy	Participant-level Accuracy
All	5	0.381	0.429	0.333	0.200	0.329	0.333
Pitch + Jitter	4	0.536	0.571	0.383	0.350	0.459	0.491
RMS-Energy	3	0.369	0.286	0.383	0.300	0.310	0.281
HNR	3	0.345	0.143	0.338	0.250	0.339	0.298
All-Prosodics	3	0.512	0.714	0.479	0.550	0.548	0.684
LSF + Δ	4	0.607	0.857	0.504	0.650	0.501	0.614
MFCC + Δ	5	0.524	0.571	0.508	0.600	0.484	0.561
CPP	5	0.485	0.561	0.508	0.550	0.548	0.571
All-Spectral	5	0.408	0.368	0.417	0.400	0.393	0.571
TEO-AM	5	1.000	1.000	1.000	1.000	1.000	1.000
TEO-FM	3	0.706	0.877	0.713	0.900	0.583	0.714
TEO-CBarea	6	0.516	0.632	0.525	0.650	0.571	0.714
TEO-RMS Energy	5	0.689	0.825	0.583	0.750	0.631	0.714
TEO-Log Energy	5	0.610	0.719	0.554	0.600	0.524	0.571
All-TEO	8	0.412	0.404	0.413	0.450	0.405	0.429
H1-H2	3	0.582	0.754	0.604	0.800	0.583	0.857
PSP	4	0.449	0.439	0.492	0.450	0.488	0.429
HRF	3	0.469	0.509	0.492	0.550	0.607	0.714
GLTP	7	0.500	0.561	0.483	0.550	0.536	0.714
All-Glottal	6	0.595	0.857	0.554	0.649	0.558	0.700

7.2 Results and Discussion

Tables 12 and 13 show the classification accuracies at the word and participant level for detecting the presence of dysarthria for each different feature types on the reduced-UA-Speech and AMSDC respectively. The word level accuracies are calculated by comparing for each word the predicted label of dysarthria or no-dysarthria to the known condition of the individual. The participant-level accuracies are the percentage of participants for which at least 50% of the individual's words were classified correctly. These two different types of accuracy calculations were both important to consider; the results report that while the word level accuracies varied between 56-75% in the reduced-UA-Speech dataset experiment, the participant level accuracies were much higher ranging from 61-92%. This finding suggests that, for the baseline models, the SVM worked very well on some individuals and very poorly on others. As one example, on the reduced-UA-Speech dataset, the prosodic features classified just 75.3% of the words correctly (0.823 precision, 0.728 recall), but 92.9% of participants were classified correctly. Participant-level classification accuracies of the prosodic feature set are compared to intelligibility score and gender in Figure 8. These results suggest a balanced model was built on the reduced-UA-Speech data that did not tend to under- or over- diagnose dysarthria, but did tend to perform poorly on certain individuals.

An important observation is the stark difference in results of the TEO-Amplitude Modulation (TEO-AM) feature set for the AMSDC experiments in Table 13 compared to the baseline experiments using only the UA-Speech dataset as shown in Table 12. While the TEO-AM feature set performed on par with the other feature sets in the baseline experiments, the cross-dataset experiments had a 100% accuracy for both the word-level and participant-level accuracy in every cross-dataset experiment. Analysis of this specific feature subset indicated the feature distribution of the test set lay completely outside of the range of the training set, a chance occurrence resulting in

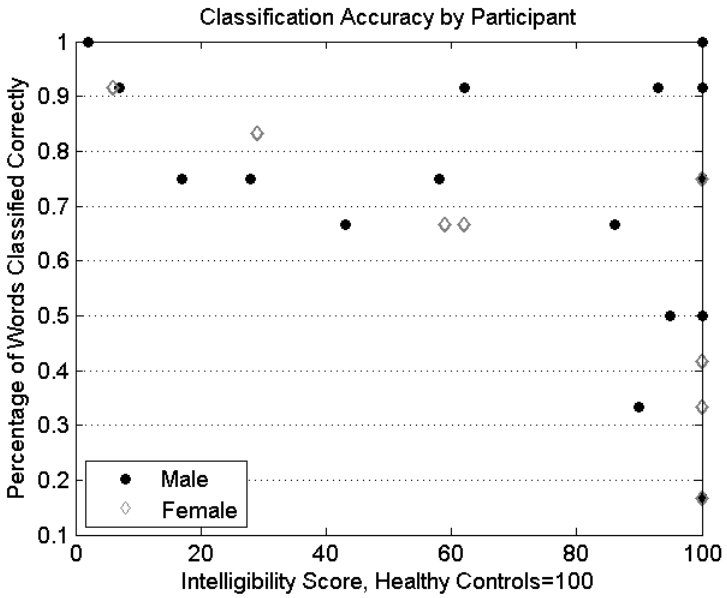

Figure 8: Dysarthria classification accuracy of each participant for prosodic feature set plotted against their intelligibility score. Plot markers differentiate gender, and healthy controls are all shown with an intelligibility score of 100.

100% accuracy based on the SVM's hyperplane due to the RBF kernel. This difference in feature distributions highlights the complexity of cross-database training in different environments, where differences in the recording environment may manifest themselves in the features calculated from the signals.

Table 13 shows the word-level accuracies of the AMSDC generally fell in the range of 30-70%, while the participant-level accuracies had a larger spread ranging from 14-85%. Additionally, many of the TEO and Glottal features were analyzed for the first time in this work with respect to there ability to detect the presence of dysarthria. The TEO-FM feature was able to achieve max word- and participant level accuracies of 0.713 and 0.9 when cross tested on the dysarthria of spastic/flaccid types on the AMSDC. This was particularly surprising as the TEO-FM was not a high performing feature for the UA-Speech training model in isolation. It was the only feature tested on the AMSDC to achieve over 0.7 in word accuracy besides the results of the TEO-AM feature subset.

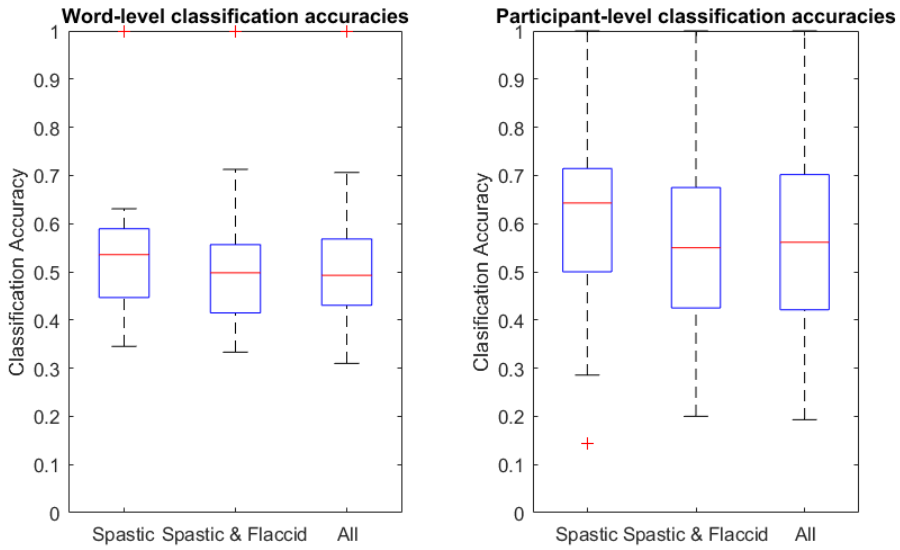

Figure 9: Cross-database analysis of the distributions of the dysarthria classification accuracies for all 20 feature subsets for each AMSDC participant subset (spastic only, spastic and flaccid only, all).

Figure 9 shows box-plots for all three AMSDC experiments summarizing the word-level (left) and participant level (right) accuracies across all 20 feature subsets for the spastic-only, spastic and flaccid, and all participant subsets. It is clear that the average accuracies are higher across all three experiments for the participant-level scores than the word-level scores, though there is no statistical significance. However, the hypothesis that the accuracies would be highest for the spastic-only participants and lowest for the full AMSDC participants was not supported by the statistics. The original hypothesis stemmed from idea that in cross-database training, the accuracies are higher when the characteristics of the individuals are matched across the two databases. In the figure, it is clear that there is no statistically significant improvement between the three box-plots for either the word or participant level. It is possible these results are manifesting from the small number of participants in the AMSDC subsets (7 spastic only, 20 spastic and flaccid), or this could stem from instead the difference in distributions of intelligibility scores between the UA-Speech dataset and

the AMSDC datasets. Since the UA-Speech dataset has a higher percentage of very-low or low intelligibility scores compared to the AMSDC, the hyperplane of the SVM model from the UA-Speech dataset may be more tuned to predict severe dysarthria in comparison to the mild dysarthria samples.

The cross-database AMSDC testing results had lower accuracies in general than those of the reduced-UA-Speech, a finding that supports the challenges of cross-database training and testing. Alghowinem et al. suggest that combining the two databases for training and using a leave-one-subject-out testing approach will result in higher accuracies due to the model being exposed to both types of speech and recording environments [123]. Future work will need to investigate this configuration for cross-database experiments and the applicability to dysarthria datasets. Additionally, previous results have reported higher error rates when attempting to separate dysarthria recordings from participants with similar types of aphasia (flaccid, spastic, and amyotrophic lateral sclerosis, 26.2%) compared to participants with types known to differ (ataxic vs amyotrophic lateral sclerosis 11.9%) [43]. The presented results support that a model trained on one type of dysarthria but tested on other types may perform poorly if the dysarthria types have differing characteristics, a trend seen in a visual inspection of Figure 9; specifically the spastic dysarthria group did have a higher average accuracy and higher second and third quartiles than the more diverse groups from the AMSDC, though the differences were not statistically significant.

To investigate the potential relationship between accuracy and intelligibility scores further, the Pearson correlation coefficients between the percentage of words correctly classified for each individual and the demographics provided for the datasets are shown in Tables 14 and 15 for the reduced-UA-Speech dataset baseline experiment and the AMSDC cross-training experiments respectively. The bold entries on Table 14 for the reduced-UA-Speech dataset show that classification accuracies for various feature sets were statistically-significantly negatively correlated with the intelligibility score.

Table 14: Correlation coefficients between participants' word-level dysarthria classification accuracies and demographic information for the reduced-UA-Speech baseline experiment. Bold entries are statistically significant, $p<0.05$ in a two-tailed test.

	Reduced UA-Speech		
Feature Type	Gender	Age	Intelligibility Score
All	-0.332	0.093	-0.368
Pitch + Jitter	-0.405	0.366	-0.388
RMS-Energy	-0.034	0.025	-0.423
HNR	0.038	0.051	**-0.524**
All-Prosodics	-0.195	-0.191	-0.223
LSF + Δ	-0.126	-0.023	**-0.553**
MFCC + Δ	0.019	0.327	**-0.787**
CPP	0.084	-0.023	**-0.533**
All-Spectral	-0.080	0.218	**-0.795**
TEO-AM	0.466	-0.005	**-0.629**
TEO-FM	0.004	0.072	**-0.570**
TEO-CBarea	-0.208	0.490	-0.280
TEO-RMS Energy	0.105	0.229	**-0.626**
TEO-Log Energy	0.136	-0.020	-0.402
All-TEO	-0.021	-0.043	-0.436
H1-H2	-0.091	0.264	**-0.756**
PSP	-0.404	0.275	-0.048
HRF	-0.039	0.333	**-0.593**
GLTP	-0.296	0.261	**-0.694**
All-Glottal	-0.304	0.081	**-0.620**

Some of these feature subsets included HNR for the prosodics, all of the spectral features, TEO-Amplitude Modulation, Frequency Modulation, and RMS-Energy, and all of the glottal features except PSP. These results suggest that the majority of feature subsets classified the best on individuals with more severe dysarthria as indicated by a lower intelligibility score. Since it is expected that those individuals with more severe dysarthria would have a larger difference in vocal acoustics and speech patterns from someone with only mild or no dysarthria, the large negative correlations values are not surprising. There were no statistically-significant correlations amongst any of the feature subsets for gender or age within the reduced-UA-Speech dataset results.

Table 15: Correlation coefficients between participants' word-level dysarthria classification accuracies and demographic information across the various AMSDC testing sets. Bold entries are statistically significant, $p<0.05$ in a two-tailed test. TEO-AM feature subset not included due to 100% accuracy causing an undefined correlation coefficient.
* indicates a moderately-high correlation, but due to n=7, not considered statistically significant.

	Spastic Only- AMSDC			Spastic and Flaccid- AMSDC			All- AMSDC		
Feature Type	Gender	Age	Intelligibility Score	Gender	Age	Intelligibility Score	Gender	Age	Intelligibility Score
All	-0.375	0.331	-0.138	0.000	0.168	-0.209	0.239	0.216	0.024
Pitch + Jitter	-0.172	-0.410	-0.137	0.214	-0.164	-0.076	**0.438**	-0.038	0.034
RMS-Energy	0.037	-0.739*	-0.167	0.263	-0.037	0.102	0.199	-0.023	0.102
HNR	0.548	0.154	-0.352	0.423	-0.109	0.137	**0.532**	0.188	-0.004
All-Prosodics	0.444	-0.289	-0.644*	**0.517**	-0.160	0.058	**0.532**	0.066	0.124
LSF + Δ	-0.018	0.496	0.106	0.025	0.132	-0.113	0.180	0.096	-0.089
MFCC + Δ	-0.016	-0.204	-0.042	-0.076	-0.140	-0.026	0.035	0.074	0.055
CPP	-0.209	0.101	-0.009	-0.068	-0.114	0.106	0.158	0.078	0.014
All-Spectral	0.151	0.213	-0.091	0.045	-0.382	0.249	**0.359**	-0.105	0.021
TEO-FM	-0.135	-0.323	-0.049	-0.218	-0.281	0.021	0.042	-0.040	-0.108
TEO-CBarea	0.057	-0.380	-0.144	0.020	-0.163	0.014	0.192	-0.026	-0.022
TEO-RMS Energy	0.061	0.837	0.360	-0.083	-0.043	-0.028	-0.008	0.023	-0.105
TEO-Log Energy	-0.139	0.470	0.285	-0.225	-0.225	-0.059	0.061	-0.007	0.016
All-TEO	0.113	0.420	-0.430	-0.037	**0.444**	-0.437	-0.059	0.057	0.004
H1-H2	-0.441	0.226	0.247	-0.157	0.368	-0.021	0.206	0.186	-0.056
PSP	0.287	-0.298	-0.465	-0.087	-0.317	-0.210	0.171	-0.149	-0.101
HRF	-0.017	0.025	-0.347	-0.052	-0.085	-0.287	0.258	0.124	-0.091
GLTP	0.288	-0.087	0.123	0.372	-0.072	0.111	**0.484**	0.085	0.069
All-Glottal	0.367	0.066	-0.173	0.434	0.111	0.072	**0.541**	0.133	0.053

As shown in Table 15, there were no statistically-significant correlations for any feature subsets applied to the AMSDC experiments and the intelligibility score. However, there was a moderately high negative correlation of the prosodic feature set to the intelligibility score in the spastic-only participants, which mimics the patterns seen in the reduced-UA-Speech dataset. This suggests that the model built on the dysarthria subtypes present in the UA-Speech dataset worked similarly when applied to the AMSDC spastic dysarthria participants, but statistical significance was not achieved with only 7 participants included in the spastic-only analysis. Additionally some of the result of the all-AMSDC analysis and the spastic and flaccid-AMSDC analysis suggests that the prosodic and glottal classification accuracies were higher for women than for men; it is unclear why the reduced-UA-Speech dataset model for dysarthria (which was trained on more males than females) would work better on the females than males in the AMSDC dataset. However, the overall lack of correlation for the accuracies in the AMSDC is likely due to training exclusively on the UA-Speech dataset for the classifier. Table 14 suggests that the intelligibility scores likely played a meaningful role in feature selection from the UA-Speech data as these served as the classifier targets in training.

This work focused on the prediction of dysarthria at the word-level recording. While the words used in training in the UA-Speech database were often multisyllabic, the words segmented from the AMSDC tended be shorter and were not segmented to be phonemically balanced. Work by Martinez et al. informally suggested shorter words were more difficult to classify into dysarthria intelligibility groupings [44]. As segmentation continues, the inclusion of more complex words from other samples in the AMSDC will be investigated to improve alignment with the UA-Speech dataset for increased classification accuracy. Additionally, the results of the AMSDC are likely influenced by the large proportion of the population presenting with a Southern African-American dialect, a characteristic not present in the UA-Speech dataset. This

study should be replicated with other datasets that would closer compare to the demographics of the UA-Speech dataset for further analysis.

CHAPTER VIII

CONCLUSIONS AND FUTURE WORK

The work presented in this dissertation analyzed the detection and prediction of stress, depression and affective state in speech with adults with aphasia through vocal acoustics analysis and machine learning. As a first step, a binary SVM classifier was applied to a subset of participants from the Aphasia Dataset in an attempt to classify depression based off of the SADQ-10 score [109]. Cepstral Peak Prominence performed the best with respect to overall accuracy, precision and recall. However, there were individuals with scores near the threshold between the binary labels with poor performance when the SVM model was applied. It has been observed that it would be hard to distinctly measure the difference of 1-2 points on the ordinal depression scales. As such, the next step was to apply a prediction model based off of regression [114]. Working with a larger subset of the Aphasia Dataset, SVR models were built to predict the SADQ-10 and PSS scores. Results were mixed, but were not precise enough to be useful as a clinical tool in the current state. This was hypothesized to be the result of short-term affective state as the prominent affective state in the vocal acoustics compared to the long-term clinical labels of stress and depression.

To test this hypothesis, changes in vocal acoustics from the first-10 and last-10 samples of each individual's recording process were compared to reported changes in affect from Self Assessment Manikins and Stress-Scale scores [117]. Statistics were computed between the distributions of features to measure potential changes in distributions and then correlated to the reported changes in affect. The result found that there was a moderate correlation between a reported change in affective score of the

stress-scale and the distribution differences. Additionally, MFCC features were the dominant feature type selected as being statistically significant to affective changes regardless of the exact feature select methodology. These finding highlight the ability to detect short-term affective state in vocal acoustics of adults with aphasia. However, they also suggest that a more complex recording and experimental process would be necessary to study long-term clinical states such as stress and depression in adults with aphasia. These findings support a longitudinal data-collection process, as well as documentation of both short- and long-term affective states at each speech recording session to have the correct emotional information and data to be able to compare changes in both short- and long-term vocal acoustics.

Throughout this entire study, the focus had been on affective state in adults with aphasia from vocal acoustics. However, the earlier work disregarded the complication that over half of the individuals in the Aphasia Database presented with at least mild dysarthria, a motor disorder that is known to impact vocal acoustics. The last component of this research was to determine the efficacy of a cross-database clinical tool to detect the presence of dysarthria [119]. The use of a cross-database study was necessary as most databases with speech from adults with aphasia tend to be either 1) limited in size, or 2) limited in scope with respect to originating medical condition or dysarthria subtype. Baseline results of training and testing on the UA-Speech dataset achieved word-level accuracies of up to 75% and participant-level accuracies of up to 93%. Once the model from the UA-Speech dataset was applied to the AMSDC, the performance dropped to 71% at the word level and 90% at the participant level. Higher accuracies were achieved when testing on the same type of dysarthria as the training model was built upon from the UA-Speech dataset. Additionally, a statistically-significant negative correlation was seen in the baseline model between speech intelligibility and classification accuracies, suggesting more-severe dysarthria with a lower intelligibility score was more likely to be classified correctly. While this

result was not seen with significance in the AMSDC results, it is possible the vocal acoustics varied more so with the subtype of dysarthria than the intelligibility score.

Future work will be necessary to create a clinical tool for automatic assessment of affective state in adults with aphasia. Preprocessing steps including detection of dysarthria presence, subtype, and severity may be necessary to ensure the changes in vocal acoustics due to dysarthria or other motor disorders are accounted for before being used in models for affective state. Additionally, a more extensive database of aphasia and affective states will be needed to continue this study. With only 20 participants available to study, very few generalizations could be made. It is recommended that careful planning take place before beginning to collect data for a new database with respect to affective states of adults with aphasia to ensure that short-term affective states are considered with respect to the long-term clinical states of interest. The work presented in this dissertation supports the ability to create a clinical tool for affective state detection and/or monitoring in adults with aphasia, but more research will be needed before it is ready for use in a clinical setting.

CHAPTER IX

CONTRIBUTIONS

The main contribution of this work is an innovative application of speech processing and machine learning to affective analysis of vocal acoustics in adults with aphasia. Prior to this work, the population with aphasia was excluded from not only technical studies from the signal processing community but also many clinical studies of post-stroke stress and depression due to their language difficulties. Focusing on a population often excluded not only allows us to recognize the importance of real data that isn't collected to make analysis easy, but also allows us to contribute new knowledge that can accompany clinical perspectives in diagnostic decisions and long-term care. The work in this study explored classification, prediction, statistical analyses that will be a starting point for future studies combining speech processing and affective analysis in adults with aphasia. As this study is a first of its kind, there are many takeaways that can enhance future work in the area, primarily the recommendation of a long-term, multi-recording setup which can be used to track vocal acoustics as well as short-term affective state changes across multiple days, weeks, or even months. Recognizing, analyzing, and compensating for the differences in short-term affective states and long-term clinical states as they impact vocal acoustics will be the major challenge of future research efforts.

REFERENCES

[1] D. Mozaffarian, E. J. Benjamin, A. S. Go, D. K. Arnett, M. J. Blaha, M. Cushman, S. R. Das, S. de Ferranti, J.-P. Desprs, H. J. Fullerton, V. J. Howard, M. D. Huffman, C. R. Isasi, M. C. Jimnez, S. E. Judd, B. M. Kissela, J. H. Lichtman, L. D. Lisabeth, S. Liu, R. H. Mackey, D. J. Magid, D. K. McGuire, E. R. Mohler, C. S. Moy, P. Muntner, M. E. Mussolino, K. Nasir, R. W. Neumar, G. Nichol, L. Palaniappan, D. K. Pandey, M. J. Reeves, C. J. Rodriguez, W. Rosamond, P. D. Sorlie, J. Stein, A. Towfighi, T. N. Turan, S. S. Virani, D. Woo, R. W. Yeh, and M. B. Turner, "Heart disease and stroke statistics2016 update," *A Report From the American Heart Association*, 2015. [Online]. Available: http://circ.ahajournals.org/content/circulationaha/early/2015/12/16/CIR.0000000000000350.full.pdf

[2] "Prevalance and most common causes of disability among adults- united states, 2005," Report, 2009. [Online]. Available: http://www.cdc.gov/mmwr/pdf/wk/mm5816.pdf

[3] N. A. Association, "Aphasia faqs," http://www.aphasia.org/aphasia-faqs/, [online; accessed type = Web Page].

[4] D. T. Wade, R. Langton Hewer, R. M. David, and P. M. Enderby, "Aphasia after stroke: Natural history and associated deficits," *Journal of Neurology, Neurosurgery and Psychiatry*, vol. 49, pp. 11–16, 1986.

[5] S. T. Engelter, M. Gostynski, S. Papa, M. Frei, C. Born, V. Ajdacic-Gross, F. Gutzwiller, and P. A. Lyrer, "Epidemiology of aphasia attributable to first ischemic stroke: Incidence, severity, fluency, etiology, and thrombolysis," *Stroke*, vol. 37, pp. 1379–1384, 2006.

[6] C. Code, G. Hemsley, and M. Haerrmann, "The emotional impact of aphasia," *Seminars in Speech and Language*, vol. 20, no. 1, pp. 19–31, 1999.

[7] L. Murray and H. Ray, "A comparison of relaxation training and syntax stimulation for chronic nonfluent aphasia," *Journal of Communication Disorders*, vol. 34, no. 1-2, pp. 87–113, 2001.

[8] J. S. Laures-Gore and L. C. DeFife, "Perceived stress and depression in left and right hemisphere post-stroke patients," *Neuropsychological Rehabilitation*, vol. 23, no. 6, pp. 783–797, 2013.

[9] S. Kouwenhoven, M. Kirkevold, K. Engedal, and H. Kim, "Depression in acute stroke: Prevalence, dominant symptoms, and associated factors. a systematic

literature review," *Disability and Rehabilitation*, vol. 33, no. 7, pp. 539–556, 2011.

[10] C. I. M. Price, R. H. Curless, and H. Rodgers, "Can stroke patients use visual analogue scales?" *Stroke*, vol. 30, pp. 1357–1361, 1999.

[11] H. E. Bennett, S. A. Thomas, R. Austen, A. M. S. Morris, and N. B. Lincoln, "Validation of screening measures for assessing mood in stroke patients," *British Journal of Clinical Psychology*, vol. 45, pp. 367–376, 2006.

[12] C. Code and M. Herrmann, "The relevance of emotional and psychosocial factors in aphasia to rehabilitation," *Neuropsychological Rehabilitation*, vol. 13, no. 1-2, pp. 109–132, 2003.

[13] M. Vidovic, O. Sinanovic, L. Sabaskic, A. Haticic, and E. Brkie, "Incidence and types of speech disorders in stroke patients," *Acta Clinica Croatia*, vol. 50, pp. 491–494, 2011.

[14] H. L. Flowers, F. L. Silver, J. Fang, E. Rochon, and R. Martino, "The incidence, co-occurence, and predictors of dsyphagia, dysarthria, and aphasia after first-ever ischemic stroke," *Journal of Communication Disorders*, vol. 46, pp. 238–248, 2013.

[15] L. Cherney and S. Small, *Aphasia, Apraxia of Speech, and Dysarthria.*, ser. Stroke Recovery and Rehabilitation. New York, NY: Demos Medical, 2009.

[16] D. F. Benson, "Fluency in aphasia: Correlation with radioactive scan localization." *Cortex*, vol. 3, pp. 373–394, 1967.

[17] M. Kerschenstiner, K. Poeck, and E. Brunner, "The fluency-nonfluency dimension in the classification of aphasic speech." *Cortex*, vol. 8, pp. 233–247, 1972. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0010945272800212

[18] A. Kertesz, *Western aphasia battery-revised (WAB-R).* Pearson, 2006.

[19] J. Ryalls, "An acoustic study of vowel production in aphasia," *Brain and Language*, vol. 29, pp. 48–67, 1986.

[20] J. Gandour, S. H. Petty, and R. Dardarananda, "Perception and production of tone in aphasia," *Brain and Language*, vol. 35, pp. 201–240, 1988.

[21] M. Vukovic, R. Sujic, M. Petrovic-Lazic, N. Miller, D. Milutinovic, S. Babac, and I. Vukovic, "Analysis of voice impariment in aphasia after stroke-underlying neuroanatomical substrates," *Brain and Language*, vol. 123, pp. 22–29, 2012.

[22] F. Goldman-Eisler, *Hesitation and information in speech*, Butterworth, London, 1961.

[23] G. Quinting, *Hesitation phenomena in adult aphasia and normal speech*, ser. Janua Linguarum. The Hauge, Netherlands: Mouton and Co., 1971.

[24] P. Enderby and R. Palmer, *Frenchay Dysarthria Assessment- Second Edition (FDA-2).* Austin, TX: PRO-ED, Inc;, 2008.

[25] M. S. Paja and T. H. Falk, "Automated dysarthria severity classification for improved objective intelligibility assessment of spastic dysarthric speech," in *13th Annual Conference of the International Speech Communication Association (INTERSPEECH-2012)*, 2012, Conference Proceedings, pp. 62–65.

[26] E. C. Guerra and D. F. Lovey, "A modern approach to dysarthria classification," in *25th Annual International Conference of the IEEE EMBS*, 2003, Conference Proceedings, pp. 2257–2260.

[27] C. Zhang, J. Dang, J. Zhang, and J. Wei, "Investigation on articulatory and acoustic characteristics of dysarthria," in *2014 9th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2014, Conference Proceedings, pp. 326–330.

[28] R. S. Tikofsky, "Phonetic characteristics of dysarthria: Final report," University of Michigan, Department of Speech and Speech Clinic, Report, 1965.

[29] K. L. Lansford and J. M. Liss, "Vowel acoustics in dysarthria: speech disorder diagnosis and classfication," *Journal of Speech, Language, Hearing Research*, vol. 57, no. 1, pp. 57–67, 2014.

[30] R. D. Kent, G. Weismer, J. F. Kent, H. K. Vorperian, and J. R. Duffy, "Acoustic studies of dysarthric speech: methods, progress, and potential," *Journal of Communication Disorders*, vol. 32, pp. 141–186, 1999.

[31] R. D. Kent and Y. J. Kim, "Toward an acoustic typology of motor speech disorders," *Clinical Linguistics and Phonetics*, vol. 17, no. 6, pp. 427–445, 2003.

[32] H. V. Sharma, "Universal access: Experiments in automatic recognition of dysarthric speech," Thesis, 2008.

[33] E. Yilmax, M. Ganzeboom, L. Beijer, C. Cucchiarini, and H. Strik, "A dutch dysarthric speech database for individualized speech therapy research," in *Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 2016, Conference Proceedings.

[34] S. R. Shahamiri and S. S. B. Salim, "A multi-views multi-learners approach towards dysarthric speech recognition using multi-nets artifical neural networks," *IEEE Trans. on Neural Systems and Rehabilitation Engineering*, vol. 22, no. 5, pp. 1053–1063, 2014.

[35] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. Huang, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research," in *9th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2008, Conference Proceedings, pp. 1741–1744.

[36] X. Menendez-Pidal, J. B. Polikoff, S. M. Peters, J. E. Leonzio, and H. T. Bunnell, "The nemours database of dysarthric speech," in *Fourth International Conference on Spoken Language (ICSLP-96)*, 1996, Conference Proceedings, pp. 1962–1965.

[37] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The torgo database of acoustic and articulatory speech from speaker wtih dysarthria," *Language Resources and Evaluation*, vol. 46, no. 4, pp. 523–541, 2012.

[38] J. R. Deller Jr., M. S. Liu, L. J. Ferrier, and P. Robichaud, "The whitaker database of dysarthric (cerebral palsy) speech," *Journal of the Acoustical Society of America*, vol. 93, no. 6, pp. 3516–3518, 1993.

[39] M. Nicolao, H. Christensen, S. Cunningham, P. Green, and T. Hain, "A framework for collecting realistic recordings of dysarthric speech- the homeservice corpus," in *10th Edition of Language Resources and Evaluation Conference (LREC 2016)*, 2016, Conference Proceedings.

[40] R. Sriranjani, M. Ramasubba Reddy, and S. Umesh, "Improved acoustic modeling for automatic dysarthric speech recognition," in *2015 Twenty First National Conference on Communications (NCC)*, 2015, Conference Proceedings.

[41] I. Laaridh, C. Fredouille, and C. Meunier, "Automatic detection of phone-based anomalies in dysathric speech," *ACM Transactions on Accessible Computing*, vol. 6, no. 3, p. 9, 2015.

[42] A. E. Aronson, "Dyarthria, differential diagnosis," 1993, [online; accessed type = Audiovisual Material].

[43] M. V. Mujumdar and R. F. Kubichek, "Design of a dysarthria classifier using global statistics of speech features," in *2010 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2010, Conference Proceedings.

[44] D. Martinez, E. Lleida, P. Green, H. Christensen, A. Ortega, and A. Miguel, "Intelligiblity assessment and speech recognizer word accuracy rate prediction for dysarthric speakers in a factor analysis subspace," *ACM Transactions on Accessible Computing*, vol. 6, no. 3, p. 10, 2015.

[45] F. Rudzicz, "Adusting dysarthric speech signals to be more intelligible," *Computer Speech & Language*, vol. 27, no. 6, pp. 1163–1177, 2013.

[46] J. Kim, N. Kumar, A. Tsiartas, M. Li, and S. S. Narayanan, "Automatic intelligibility classification of sentence-level pathological speech," *Computer Speech & Language*, vol. 29, no. 1, pp. 132–144, 2015.

[47] J. Carmichael, "Dysarthria diagnosis via respiration and phonation," in *6th International Conference and Workshop on Computing and Communication (IEMCON)*, 2015, Conference Proceedings, pp. 1–5.

[48] A. DeMinio, R. F. Kubichek, and K. Caves, "Assessing dysarthria severity using global statistics and boosting," in *2011 Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems, and Computers (ASILOMAR)*, 2011, Conference Proceedings, pp. 1103–1107.

[49] P. Square-Storer and E. A. Roy, *The Apraxias: Commonalities and Distinctions*, ser. Acquired Apraxia of Speech in Aphasic Adults. Hillsdale: Lawrence Erlbaum Associates, Publishers, 1989.

[50] F. L. Darley, "Apraxia of speech: 107 years of terminological confusion," *Paper presented to the American Speech and Hearing Association, Denver Colorado (unpublished)*, 1968.

[51] B. L. Dabul, *Apraxia Battery for Adults, Second Edition.* Austin, TX: PRO-ED, 2000.

[52] J. Ogar, H. Slama, N. Dronkers, S. Amici, and M. L. Gorno-Tempini, "Apraxia of speech: an overview," *Neurocase*, vol. 11, pp. 427–432, 2005.

[53] R. W. Picard, *Affective Computing.* Cambridge, Massachusetts: MIT Press, 1997.

[54] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, pp. 32–80, 2001.

[55] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, "Emotional speech: Towards a new generation of databases," *Speech Communication*, vol. 40, pp. 33–60, 2003.

[56] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Communication*, vol. 48, pp. 1162–1181, 2006.

[57] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[58] G. Zhou, J. H. L. Hansen, and J. F. Kaiser, "Nonlinear feature based classification of speech under stress," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 201–216, 2001.

[59] H. Kurniawan, A. V. Maslov, and M. Pechenizkiy, "Stress detection of speech and galvanic skin response signals," in *2014 IEEE 27th International Symposium on Computer-Based Medical Systems*, 2013, Conference Proceedings, pp. 209–214.

[60] B. D. Womack and J. H. L. Hansen, "Stress independent robust hmm speech recognition using neural network stress classification," in *4th European Conference on Speech, Communication, and Technology (EUROSPEECH '95')*, 1995, Conference Proceedings, pp. 1999–2002.

[61] H. Ellgring and K. R. Scherer, "Vocal indicators of mood change in depression," *Journal of Nonverbal Behavior*, vol. 20, no. 2, pp. 83–110, 1996.

[62] A. Ozdas, R. G. Shiavi, S. E. Silverman, M. K. Silverman, and D. M. Wilkes, "Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 9, pp. 1530–1540, 2004.

[63] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jian, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, "Avec 2013- the continuous audio/visual emotion and depression recognition challenge," in *AVEC'13 Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*. ACM, 2013, Conference Proceedings, pp. 3–10.

[64] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, and R. Cowie, "Avec 2016- depression, mood, and emotion recognition workshop and challenge," in *Audio/Visual Emotion Challenge and Workshop (AVEC 2016)*, 2016, Conference Proceedings, pp. 3–10.

[65] N. Cummins, S. Scherer, J. Krajewski, S. Schneider, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, 2015.

[66] D. Le, K. Licata, C. Persad, and E. Mower Provost, "Automatic assessment of speech intelligibility for individuals with aphasia," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2187–2199, 2016.

[67] T. Mabutchi, T. Takeda, and T. Matuda, "Aphasia rehabilitation support system by using multimodal interface device," in *2015 IEEE Conference on Systems, Man, and Cybernetics*, 2015, Conference Proceedings, pp. 1433–1438.

[68] H. Goodglass, E. Kaplan, and B. Barresi, "Boston diagnostic aphasia examination- third edition (bdae-3)," 2000, [online; accessed type = Generic].

[69] L. E. Nicholas and R. H. Brookshire, "A system for quantifying the informativeness and efficiency of the connected speech of adults with aphasia," *Journal of Speech, Language, and Hearing Research*, vol. 36, pp. 338–350, 1993.

[70] L. M. Sutcliffe and N. B. Lincoln, "The assessment of depression in aphasic stroke patients: the development of the stroke aphasic depression questionaire," *Clinical Rehablitation*, vol. 12, no. 6, pp. 506–513, 1998.

[71] C. Benaim, B. Cailly, D. Perennou, and J. Pelissier, "Validation of the aphasic depression rating scale," *Stroke*, vol. 40, no. 2, pp. 523–529, 2004.

[72] L. Leeds, R. Meara, and J. Hobson, "The utility of the stroke aphasia depression questionaire (sadq) in a stroke rehabilitation unit," *Clinical Rehablitation*, vol. 18, pp. 228–231, 2004.

[73] J. Laures-Gore, M. Farina, E. Moore, and S. Russell, "Stress and depression scales in aphasia: relation between the aphasia depression rating scale, stroke aphasia depression questionnaire-10, and the perceived stress scale," *Topics in Stroke Rehabilitation*, pp. 1–5, 2016. [Online]. Available: http://dx.doi.org/10.1080/10749357.2016.1198528

[74] S. Cohen, T. Kamarck, and R. Mermelstein, "A global measure of perceived stress," *Journal of Health and Social Behavior*, vol. 24, no. 4, pp. 385–396, 1983.

[75] N. Y. S. U. T. Union, "Stress assessments: Perceived stress scale," http://www.nysut.org/~/media/files/nysut/resources/2013/april/social-services/socialservices_stressassessments.pdf?la=en, [online; accessed type = Web Page].

[76] N. H. D. of Administrative Services, "Perceived stress scale," "http://das.nh.gov/wellness/Docs%5CPercieved%20Stress%20Scale.pdf", [online; accessed type = Web Page].

[77] M. M. Bradley and P. J. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," *Journal of Behavior Therapy and Experimental Psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[78] J. Laures-Gore, C. M. Heim, and Y.-S. Hsu, "Assessing cortisol reactivity to a lingustic task as a marker of stress in individuals with left-hemisphere stroke and aphasia," *Journal of Speech, Language, and Hearing Research*, vol. 50, pp. 493–507, 2007.

[79] J. S. Laures-Gore, S. Russell, R. Patel, and M. Frankel, "The atlanta motor speech disorders corpus: Motivation, development, and utility," *Folia Phoniatrica et Logopaedica*, vol. 68, no. 2, pp. 99–105, 2016.

[80] R. Patel, K. Connaghan, D. Franco, E. Edsall, D. Forgit, L. Olsen, L. Ramage, E. Tyler, and S. Russell, ""the caterpillar": A novel reading passage for assessment of motor speech disorders," *American Journal of Speech-Language Pathology*, vol. 22, pp. 1–9, 2013.

[81] C. Van Riper, *Speech correction: Principles and methods*, 4th ed. Englewood Cliffs, NJ: Prentice-Hall, 1963.

[82] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of ACM Multimedia (MM)*. ACM, 2013, Conference Proceedings, pp. 835–838.

[83] A. Alwan, "Voicesauce: A program for voice analysis," http://www.ee.ucla.edu/spapl/voicesauce/, 2012, [online; accessed type = Web Page].

[84] J. Hillenbrand, R. A. Cleveland, and R. L. Erickson, "Acoustic correlates of breathy voice," *Journal of Speech and Hearing Research*, vol. 37, pp. 769–778, 1994.

[85] Y. Maryna, N. Roy, M. De Bodt, P. Van Cauwenberge, and P. Corthals, "Acoustic measurement of overall voice quality: A meta-analysis," *Journal of the Acoustical Society of America*, vol. 126, no. 5, pp. 2619–2634, 2009.

[86] P. J. Murphy, "Periodicity estimation in synthesized phonation signals using cepstral rahmonic peaks," *Speech Communication*, vol. 48, pp. 1704–1713, 2006.

[87] G. de Krom, "A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals," *Journal of Speech and Hearing Research*, vol. 36, no. 2, pp. 254–266, 1993.

[88] S. Deb and S. Dandapat, "A novel breathiness feature for analysis and classification of speech under stress," in *IEEE Twenty First National Conference On Communications*, 2015, Conference Proceedings.

[89] M. Brookes, "Voicebox: Speech processing toolkit for matlab," http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html, 2014, [online; accessed type = Web Page].

[90] J. R. Orozco-Arroyave, F. Honig, J. D. Arias-Londono, J. F. Vargas-Bonilla, K. Daqrouq, S. Skodda, J. Rusz, and E. Noth, "Automatic detection of parkinson's disease in running speech spoken in three different languages," *Journal of the Acoustical Society of America*, vol. 139, no. 1, pp. 481–500, 2016.

[91] J. Rusz, R. Cmejla, H. Ruzickova, and E. Ruzicka, "Quantitiative acoustic measurements for characterization of speech and voice disorders in early untreated parkinson's disease," *Journal of the Acoustical Society of America*, vol. 129, no. 1, pp. 350–367, 2011.

[92] B. Hu, Z. Liu, L. Yan, T. Wang, F. Liu, Z. Li, and H. Kang, "Feature selection and classification of speech under long-term stress," in *2015 IEEE International Conference on Bioinformatics and Biomedicine*, 2015, Conference Proceedings, pp. 904–910.

[93] L.-S. A. Low, N. C. Maddage, M. Lech, and N. Allen, "Mel frequency cepstral feature and gaussian mixtures for modeling clinical depression in adolescents," in *8th IEEE International Conference on Cognitive Informatics*, 2009, Conference Proceedings, pp. 346–350.

[94] L. R. Rabiner and R. W. Schafer, *Theory and Applications of Digital Speech Processing*, first edition ed. Upper Saddle River, NJ: Pearson, 2011.

[95] X. Sun, "Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio," in *ICASSP 2002*, 2002, Conference Proceedings, pp. 333–336.

[96] J. B. Alonso, J. Cabrera, M. Medina, and C. M. Travieso, "New approach in quantification of emotional intensity from the speech signal: emotional temperature," *Expert Systems With Applications*, vol. 42, pp. 9554–9564, 2015.

[97] T. Banziger and K. R. Scherer, "The role of intonation in emotional expressions," *Speech Communication*, vol. 46, pp. 252–267, 2005.

[98] E. Bozkurt, E. Erzin, C. Eroglu Erdem, and A. Tanju Erdem, "Use of line spectral frequencies for emotion recognition from speech," in *2010 IEEE International Conference on Pattern Recognition*, 2010, Conference Proceedings, pp. 3708–3711.

[99] M. Suzuki, S. Nakagawa, and K. Kita, "Emotion recognition method based on normalization of prosodic features," in *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2013, Conference Proceedings.

[100] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, G. Parker, and M. Breakspear, "Characterising depressed speech for classification," in *14th Annual conference of the International Speech Communication Association (InterSpeech)*, 2013, Conference Proceedings, pp. 2534–2538.

[101] L.-S. A. Low, N. C. Maddage, M. Lech, L. B. Sheeber, and N. B. Allen, "Detection of clinical depression in adolescents' speech during family interactions," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 3, pp. 574–586, 2011.

[102] E. Moore II, M. Clements, J. Peifer, and L. Weisser, "Investigating the role of glottal features in classifying clinical depression," in *Proceedings of the 25th Annual International Conference of the IEEE-EMBS*, 2003, Conference Proceedings, pp. 2849–2852.

[103] ——, "Comparing objective feature statistics of speech for classifying clinical depression," in *Proceedings of the 26th Annual International Conference of the IEEE-EMBS*, 2004, Conference Proceedings, pp. 17–20.

[104] ——, "Critical analysis of the impact of glottal features in the classification of clinical depression of speech," *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 1, pp. 96–1070, 2008.

[105] I. R. Titze and J. Sundberg, "Vocal intensity in speakers and singers," *Journal of the Acoustical Society of America*, vol. 91, no. 5, pp. 2936–2946, 1992.

[106] D. Childers and C. Lee, "Vocal quality factors: Analysis, synthesis, and perception," *Journal of the Acoustical Society of America*, vol. 90, no. 5, pp. 2394–2410, 1991.

[107] P. Alku, H. Strik, and E. Vilkman, "Parabolic spectral paramater- a new method for quanitification of the glottal flow," *Speech Communication*, vol. 22, pp. 67–79, 1997.

[108] J. F. Torres, E. Moore II, and E. Bryant, "A study of glottal waveform features for deceptive speech classification," in *2008 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2008, Conference Proceedings, pp. 4489–4492.

[109] S. Gillespie, E. Moore II, J. S. Laures-Gore, and M. Farina, "Exploratory analysis of speech features related to depression in adults with aphasia," in *41st IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2016, Conference Proceedings, pp. 5185–5189.

[110] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *10th Annual Conference of the International Speech Communication Association*, 2009, Conference Proceedings, pp. 312–315.

[111] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Peutemann, and I. H. Witten, "The weka data mining software: An update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[112] M. A. Hall, "Correlation-based feature selection for machine learning," Thesis, 1999.

[113] R. Fraile and J. I. Godino-Llorente, "Cepstral peak prominence: A comprehensive analysis," *Biomedical Signal Processing and Control*, vol. 14, pp. 42–45, 2014.

[114] S. Gillespie, E. Moore II, J. Laures-Gore, M. Farina, S. Russell, and Y.-Y. Logan, "Detecting stress and depression in adults with aphasia through speech analysis," in *42nd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2017)*, 2017, Conference Proceedings.

[115] N. Wahidah Hashim, M. Wilkes, R. Salomon, J. Meggs, and D. J. France, "Evaluation of voice acoustics as predictors of clinical depression scores," *Journal of Voice*, vol. In Press, 2016.

[116] M. Hamilton, "A rating scale for depression," *Journal of Neurology, Neurosurgery and Psychiatry*, vol. 23, pp. 56–62, 1960.

[117] S. Gillespie, J. Laures-Gore, E. Moore II, M. Farina, and S. Russell, "Identification of affective state in adults with aphasia using speech acoustics," *Journal of Speech, Language, and Hearing Research*, 2017- Submitted.

[118] J. D. Evans, *Straightforward Statistics for the Behavioral Sciences.* Pacific Grove, CA: Brooks/Cole Publishing, 1996.

[119] S. Gillespie, Y.-Y. Logan, E. Moore, J. Laures-Gore, R. Patel, and S. Russell, "Cross-database models for the classification of dysarthria presence," in *Annual Conference of the International Speech Communication Association (INTERSPEECH 2017)*, 2017- Submitted, Conference Proceedings.

[120] J. Mekyska, Z. Smekal, Z. Galaz, Z. Mzourek, I. Rektorova, M. Faundez-Zanuy, and K. Lpez-de Ipia, *Perceptual Features as Markers of Parkinsons Disease: The Issue of Clinical Interpretability.* Cham: Springer International Publishing, 2016, pp. 83–91.

[121] T. M. DeCicco and R. Patel, "Machine classification of prosodic control in dysarthria," *Journal of Medical Speech-Language Pathology*, vol. 18, no. 4, pp. 35–39, 2010.

[122] G. Vyas, M. K. Dutta, J. Prinosil, and P. Harar, "An automatic diagnosis and assessment of dysarthric speech using speech disorder specific prosodic features," in *2016 29th International Conference on Telecommunications and Signal Processing (TSP)*, 2016, Conference Proceedings, pp. 515–518.

[123] S. Alghowinem, R. Goecke, J. Epps, M. Wagner, and J. Cohn, "Cross-cultural depression recognition from vocal biomarkers," in *Conference of the International Speech Communication Association (Interspeech-2016)*, 2016, Conference Proceedings.

[124] M. Tahon, M. A. Sehili, and L. Devillers, *Cross-Corpus Experiments on Laughter and Emotion Detection in HRI with Elderly People.* Cham: Springer International Publishing, 2015, pp. 633–642.