Characterization of Phoneme Rate as a Vocal Biomarker of Depression

by

Gregory Alan Ciccarelli

B.S., Electrical Engineering, The Pennsylvania State University, 2009
S.M., Electrical Engineering and Computer Science, M.I.T., 2013

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in Electrical Engineering and Computer Science
at the Massachusetts Institute of Technology
June 2017

Signature of Author: Signature redacted
Department of Electrical Engineering and Computer Science
May 19, 2017

Certified by: Signature redacted
John D. E. Gabrieli
Grover Hermann Professor in Health Sciences and Technology
and Cognitive Neuroscience
Thesis Supervisor

Certified by: Signature redacted
Satrajit S. Ghosh
Principal Research Scientist, the McGovern Institute for Brain Research
Thesis Supervisor

Certified by: Signature redacted
Thomas F. Quatieri
Senior Technical Staff, MIT Lincoln Laboratory
Thesis Supervisor

Accepted by: Signature redacted
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Committee for Graduate Students

Characterization of Phoneme Rate as a Vocal Biomarker of Depression

by Gregory Alan Ciccarelli

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Abstract

Quantitative approaches to psychiatric assessment beyond the qualitative descriptors in the Diagnostic and Statistical Manual of Mental Disorders could transform mental health care. However, objective neurocognitive state estimation and tracking demands robust, scalable indicators of a disorder. A person's speech is a rich source of neurocognitive information because speech production is a complex sensorimotor task that draws upon many cortical and subcortical regions. Furthermore, the ease of collection makes speech a practical, scalable candidate for assessment of mental health. One aspect of speech production that has shown sensitivity to neuropsychological disorders is phoneme rate, the rate at which individual consonants and vowels are spoken. Our aim in this thesis is to characterize phoneme rate as an indicator of depression and to improve our use of phoneme rate as a feature through both brain imaging and neurocomputational modeling.

This thesis proposes that psychiatric assessment can be enhanced using a neurocomputational model of speech motor control to estimate unobserved parameters as latent descriptors of a disorder. We use depression as our model disorder and focus on motor control of speech phoneme rate. First, we investigate the neural basis for phoneme rate modulation in healthy subjects uttering emotional sentences and in depression using functional magnetic resonance imaging. Then, we develop a computational model of phoneme rate to estimate subject-specific parameters that correlate with individual phoneme rate. Finally, we apply these and other features derived from speech to distinguish depressed from healthy control subjects.

Thesis Supervisor: John D. E. Gabrieli
Title: Grover Hermann Professor in Health Sciences and Technology
and Cognitive Neuroscience

Thesis Supervisor: Satrajit S. Ghosh
Title: Principal Research Scientist

Thesis Supervisor: Thomas F. Quatieri
Title: Senior Technical Staff

Acknowledgments

In grateful recognition of the individuals and groups who have invested themselves in my doctoral journey, I echo Carlo's often repeated exclamation, with emphasis on the plural pronoun, "We've got this!"

- Thomas Quatieri, my Lincoln co-supervisor, for his attention to detail in communication and his generosity with his time.
- Satrajit Ghosh, my campus co-supervisor, for his patience and his views on the scientific process.
- John Gabrieli, my campus co-supervisor, for his thoughtful feedback and his acceptance of an EECS student into his cognitive neuroscience laboratory.
- Thomas Heldt and Polina Golland, for their guidance as members of my committee.
- The Gabrieli Laboratory and Voice project contributors with special mention of Kevin Sitek, Carlo de los Angeles, Mathias Goncalves, Anissa Sridhar, and our experiment participants.
- The SHBT and BCS entering classes of 2014, and the Fee, Golland, and McDermott labs for their friendship and collegiality.
- Jeffrey Palmer and the Lincoln Scholars committee for this opportunity.
- My colleagues at Lincoln with special mention of Adam Lammert and Chris Smalt for productive discussions and the Lincoln librarians for their helpfulness.
- My extended family, my brother, Geoff, and especially my parents, Victoria and Dennis, who have supported me throughout my career with a saint's level of patience, support, and love.
- All the saints and members of heaven who have interceded on my behalf, and God for carrying me through to the end.

This work was supported by an MIT Lincoln Laboratory scholarship, by an MIT Lincoln Laboratory and a McGovern Institute for Neurotechnology Grant, and an NIH NIBIB R01 EB020740. This work used the Martinos Imaging Center at MIT. Parkinson's data was contributed by users of the Parkinson's mPower mobile application as part of the mPower study developed by Sage Bionetworks and described in Synapse doi:10.7303/syn4993293.

Contents

List of Figures

Chapter 1

Introduction

THIS thesis investigates the neurobiology of depression's effect on speech through neurocomputational models with the aim of advancing the practical application and scientific understanding of using speech as a biomarker of depression. We contribute a novel analysis paradigm that uses neurocomputational models of speech to identify biomarkers of depression, and we investigate depression's effect on a specific aspect of speech production, phoneme rate control.

1.1 Need for Depression Biomarkers

A major depressive episode afflicts 6.7% of the adult US population each year [25], and major depression disorder costs the US $210 billion annually [58]. A key step to combating depression effectively is to develop biomarkers for timely detection of depression and for tracking its severity in the presence or absence of treatment. A biomarker of depression is a datum derived from a measurement of an individual that relates to the individual's depression severity. This information may be helpful to caregivers and also to the individual by making the individual aware of the condition [73].

Biomarkers are needed because methods for assessing depression are unreliable, slow to administer, subjective, late to diagnose the problem, and only intermittently performed. The gold standard by which depression is diagnosed is the structured clinical interview (SCI) by a trained clinician. The SCI is used to evaluate a subject for essentially qualitative, not quantitative, signs of depression where the signs of depression are set forth in the Diagnostic and Statistical Manual of Mental Disorders (DSM) [5]. To be diagnosed with major depressive disorder (MDD), the DSM V [4] requires five of nine conditions to be met, and at least one of the five conditions must be either depressed mood or loss of pleasure in life. Unfortunately, even though trained clinicians are the only ones certified to make the diagnosis, even among them there is large variability. Diagnosing MDD has a low inter-rater agreement with a kappa value of 0.28 (compare to schizophrenia with kappa of 0.46 and post traumatic stress disorder with kappa of 0.67 [109]).

Because the SCI is limited to clinicians as well as time consuming to administer (it can take between 45 minutes to two hours), alternative patient self-report surveys

have been created. There are a plethora of both open source and proprietary measures. These include the Beck Depression Inventory [10], the Quick Inventory of Depression [112], and the Patient Health Questionnaire [75]. The existence of so many screening tools highlights the challenge of identifying depression, but also presents obstacles to progress as different studies might use different tools. Therefore results may not be directly comparable.

We highlight one self-report survey, the Beck Depression Inventory, for its use in this thesis. The BDI was originally published in 1961 [9] and has since been revised. The current version, and the one used in this thesis, is the BDI-II [10], though for brevity the second edition version will be abbreviated as BDI. The BDI is a 21 question, multiple choice survey with a minimum score of 0 and a maximum score of 63. In this thesis, a cutoff score of greater than or equal to 14 is used to classify a subject as depressed according to the guidelines established by Beck et al. [10].

One major limitation of all predictive evaluations in this thesis is that our predictions can be no better than the accuracy of the BDI reported by the subject. We must acknowledge this limitation, but we also see using the BDI for validation as a means to build credibility in objective measures. The long term goal would be to use biomarkers both to predict onset of depression and individual-appropriate intervention.

Other problems with the current clinician centered system are the delays before diagnosis is made, the availability of clinicians, and the infrequency of follow up. Huerta-Ramírez et al. [69] found an average delay of 9.89 weeks from time of symptom onset in a large study in Spain, and in Belgium only 14 percent of people sought treatment within a year of symptom onset[20]. With the diversity of qualitative measures and the reliance on variable professional opinion as the gold standard, there is a clear need for an objective, low cost, fast alternative for tracking depression.

■ 1.2 Motivation and the Phoneme Rate Biomarker

Biomarkers derived from speech have shown predictive value in associating an individual with a depression level (see Cummins et al. [33] for a review). Speech in particular is well suited as a source of biomarkers for mental illnesses because of its ease of collection, and potential to scale globally by taking advantage of ubiquitous, wireless mobile devices (e.g. smart-phones). This thesis focuses on low level acoustic phenomenon, but other levels of communication analysis exist (e.g. language content and communication behaviors [37]), and should be considered in a full voice and language based assessment system. The array of candidate features that may be extracted from the voice include measures of voice quality, pitch stability and variance, clarity of speech, fluency, grammatical complexity, and level of interactivity with another person.

A person's overall speaking rate was one of the earliest investigated voice based biomarkers of depression because of its ease of measurement and hypothesized sign of psychomotor retardation. However, overall speaking rate has had mixed success as a depression biomarker. Neither Darby and Hollien [34] nor Nilsonne [95] found speak-

ing rate useful, but Ellgring and Scherer [40] found speech rate increased in subjects recovering from depression. A review by Sobin and Sackeim [123] did find that depressed subjects had greater speech pause time, which would slow overall speech rate. Consistent with Sobin and Sackeim [123], Mundt et al. [93] found depression patients responding to treatment increased speech rate over baseline.

Trevino et al. [129] advanced the field by moving from gross measures of rate to fine grained measures of timing. Specifically, they analyzed the timing of individual units of speech, consonants or vowels. These individual units of speech sound or phonemes are spoken with different average durations. Collectively the timing differences have been called the phoneme rate biomarker because the average speed at which a person speaks a particular phoneme within their words and sentences is called a phoneme rate. Williamson et al. [144] then used the phoneme rate biomarker as part of the best depression prediction algorithm in an international competition for predicting depression severity.

Despite the phoneme rate biomarker's success, there is little known about the neural mechanism by which phoneme rate control would be tied to depression severity. We address this open question through a functional magnetic resonance (fMRI) investigation of overtly produced, read, implicitly emotional speech in a depressed population. We aim to encapsulate the science behind this biomarker within a computational modeling framework. We believe that doing so may lead to features from the model that improve depression tracking. Ultimately, knowledge encapsulated in a model may help with treatment decisions.

■ 1.3 This Thesis as a Framework

We propose this thesis as a framework in which to advance tracking of neuropsychological disorders through computational models. The approach of this thesis is an example of how different disciplines can be profitably unified to understand the science behind phenomenology and summarize findings in ways that can be practically used. We bring together four components and use them together to understand the science and practical application of the phoneme rate biomarker. These components in general are the brain, a neuropsychiatric disorder, computational modeling, and quantitative biomarkers. This thesis makes these components concrete by focusing on depression, neurocomputational models of speech production, and the phoneme rate biomarker. However, the general principle of focusing on observed phenomenology and using a brain inspired computational model to identify it can be applied to other disorders.

■ 1.3.1 Why Create Computational Models?

We stress that modeling is the glue that binds together observations and neurobiological knowledge. A computational model is a mathematically defined set of inputs, outputs, and the functions or algorithms that relate inputs and outputs. A neurocomputational model is a computational model whose inputs, outputs, and especially functions, are

constrained or inspired by neurobiological principles.

Neurocomputational models have several strengths. First, a neurocomputational model acts as a human interpretable summary of knowledge about the process under study. Such a summary demands clarity of thought about the process which facilitates additional knowledge acquisition and communication of understanding to to others. A model makes transparent the pieces of a system and the level of complexity at which the pieces are developed as well as the assumptions made about the system's rules. Consequently, models can point towards future experiments that should be taken as well as reveal characteristics of the system under study that are not in the data that was collected [98].

Second, with a summary of how a disorder interacts with different mechanisms of the rest of the system, researchers may be able to identify therapeutic interventions or make predictions about outcomes of interventions. This is a domain in which computational models are advantageous relative to qualitative models. Specifically, a neurocomputational model allows *in silico* experiments. Some experiments may be too costly, dangerous, or otherwise experimentally difficult to implement, but a model allows hundreds of variants of an experiment to be explored with no risk. With quantitative predictions about experimental interventions, models provide a direct means of comparing decisions. Computational models also allow reconciling quantitative experimental data with model predictions in order to refine model structure [98].

Third, a neurocomputational model enables inference from symptoms to mechanisms of symptoms. This is the particular practical use to which we apply neurocomputational models in this thesis. A neurocomputational model may permit an inverse solution in which observations are linked back to underlying causes. Importantly, the inverse solution of cause from effect may be ill-posed without using the constraints imposed by a neurocomputational model. A neurocomputational model's inherent constraints act as a prior on what can be a possible causative mechanism. Consequently, decisions based on inference through models may be robust to noise and artifacts, and invariant under perturbations to the raw biomarkers themselves [98].

■ 1.3.2 Computational Models in Psychiatry and Speech Production

In the field of computational psychiatry, others have proposed the use of computational models to identify neuropsychological disorders [90, 140]. The general approach to inverse modeling the brain through observed behavior is outlined by Wiecki et al. [140],

> Parameters of a computational model are fit to a subject's behavior on a suitable task or task battery. Different parameter values point to differences in underlying neurocircuitry of the associated subject or subject group. These parameters can be used either comparatively to study group differences (e.g., healthy and diseased) or as a regressor with, for example, symptom severity.

Despite substantial development of several neurocomputational models of speech

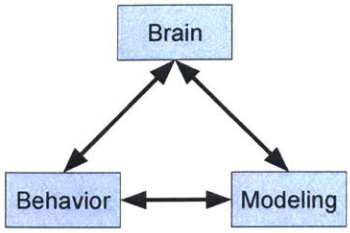


Figure 1.1: This thesis brings together three key elements and studies their relationships. These elements are neurobiology ("Brain"), behavioral responses, in particular speech ("Behavior"), and computational modeling ("Modeling").

production [61, 68, 76], and extensive analysis of the speech waveform to identify neuropsychological disorders in the decades since some of the earliest publications (Parkinson's disease, Canter [23]; depression, Darby and Hollien [34]), using neurocomputational models of speech to identify disorders is virtually unknown in the speech analysis literature. Gómez-Vilda et al. [54] recognized that disorders with a neurological basis such as Parkinson's can impact speech output, and that parameters from a biomechanical model of the speech system could be informative. However, the brain is at the root of neuropsychological disorders, so we propose that computational models should emphasize the neural control of the speech production process and not be limited to a biomechanical model of the vocal source alone.

■ 1.3.3 Research Approach of this Thesis

This thesis is a specific instance of a research framework that aims to understand the mechanisms of a neuropsychological disorder and the relationship of that disorder to observed symptoms through computational models. Figure 1.1 graphically depicts the three key components: the brain, a person's behavior as measured through his speech, and computational models that seek to unify neural and behavioral responses.

We used functional magnetic resonance imaging (fMRI) experiments to attempt to gain some neurobiological understanding that can be used to inform the model, and by using the model itself, we attempt to perform inference on the unobserved latent state of the person (e.g., the person's depression severity). In addition, providing new insight into methods of sensorimotor integration with cognitive processes may lead to better identification, tracking, and prediction of treatment efficacy not just for depression but also Parkinson's, Alzheimer's, traumatic brain injury (TBI), and autism, as examples.

The brain, in all its complexity, is a difficult entity to understand. The brain can be studied at multiple levels, and this thesis approaches the brain at a system level. As a nucleating paradigm around which we could fashion a research program, we narrowly focused on one component of speech, phoneme rate, and one particular neuropyschological disorder, depression. This narrow scope necessarily limited our explorations,

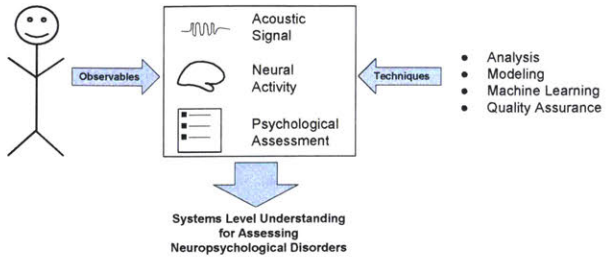


Figure 1.2: This thesis aims to understand an individual at a systems level by applying analysis, modeling, machine learning, and quality assurance techniques to acoustic and neuroimaging observeables of psychological health.

but it did provide structure for our investigative techniques of analysis, modeling, and machine learning.

As shown in Figure 1.2, an individual is a complex system from which we can gather multiple (noisy) measurements. We collected an assessment of psychological health through a self-reported depression inventory, neural activity through task-based fMRI, and speech production function through acoustics. Then, we sought to relate our available measurements to psychological health both directly and through models that attempted to capture operational mechanisms within these domains. A particular challenge to this thesis because of its system level approach is the complexity of the big picture. From data collect to processing, we must be cognizant of the input data quality and the appropriateness of the processing algorithms to be confident our results are meaningful. While any individual modality could have been developed in-depth, we chose a system level approach as an opportunity to relate different disciplines in order to use speech to track neuropyschological disorders.

■ 1.4 Summary of Contributions

This thesis makes contributions to data collection, analysis, and modeling that support our conceptual advance: the introduction of neurocomputational models for neurological disorder assessment to the speech analysis community. We advocate this method as a powerful, general research framework for unifying investigations across disorders and experiments within a quantitative, model-based framework. We first proposed this idea in Williamson et al. [145] as part of a larger effort to assess Parkinson's severity, and have since developed this idea as discussed below with a particular focus on phoneme rate and depression.

Established a multi-faceted MRI and voice data collection

We establish a multi-faceted dataset of a variety of tasks in a depressed and control population. Using the state-of-the-art fMRI acquisition technique of simultaneous multi-

slice acquisition, we collect eight task-based functional MRI runs as well as a resting scan, a diffusion weighted imaging scan, and structural scans. A defining feature of our protocol was the inclusion of six overt speech production tasks intended to elucidate how depression interacts with the brain during different aspects of speech production. These aspects included speech rate, prosody (the rhythm and timing of speech), and sequencing (how vowels and consonants are synthesized into syllables and words).

Characterized the phoneme rate biomarker and its neural basis

We perform a replication study to assess the viability of the phoneme rate biomarker, and go beyond previous studies in this area to assess the generalizability of the phoneme rate biomarker across different read passages. We analyze fMRI and acoustic data from one of the fMRI tasks, overt production of implicitly emotional sentences. We study the interaction of depression, phoneme rate, and stimulus valence to advance our scientific understanding of the phoneme rate biomarker with a view towards using this knowledge to build a computational model to aid in tracking depression. We achieve a receiver operating characteristic area under the curve (ROC AUC) of 0.73.

Created models of vocal source control and phoneme rate control, and a predictive model of depression severity from speech

We have applied the neural control framework to the vocal source. To do so, we developed an approximation to the vocal source as a new element to the computational model of speech production. This approximation focuses on the neural control of the biophysical model of the vocal source. Different than Larson et al. [79], our full neural model of source control includes both this biophysical source model, and the sensorimotor transformation between acoustic and motor space. We then applied this vocal source model to both depression and Parkinson disorder with some promising results [29]. We develop a model of phoneme rate variability that integrates phonemes, rate, and sensorimotor processing with an algorithm that allows fitting the model to an individual's speech.

We compare the features derived from model-based and model-free features. We answer the question of how model-based features compare to model-free features and consider whether model-based features add value when combined with model free features. We find peak ROC AUC performance on model-free features of 0.82, and peak performance using phoneme rate model features of 0.55.

■ 1.5 Thesis Organization

Chapter 2 presents an analysis of the phoneme rate biomarker on read speech. Chapter 3 moves into the neuroscience facet of the thesis with an fMRI investigation of the phoneme rate biomarker. Chapter 4 creates a neurocomputational model for the phoneme rate biomarker. Chapter 5 takes this thesis's work in neurocomputational modeling and the phoneme rate biomarker into a full-featured, voice-based, depression

analysis system. It compares model-based and model-free vocal biomarkers of depression. Chapter 6 summarizes the thesis's contributions and highlights promising research directions.

Chapter 2

Characterization of Phoneme Rate in Read Speech

IN this chapter, we consider two issues that have practical bearing on how the phoneme rate biomarker might be applied in real applications. First, we consider how passage duration affects the ability to discriminate between depressed and control subjects. How many seconds of speech are needed for accurate assessment of depression severity? How does accuracy of assessment relate to number of seconds of analyzed speech? Second, we consider how stable phoneme rate features are across different read passages. Does an assessment tool designed using one speech passage generalize across passages?

Section 2.1 formally defines phoneme rate and related terminology. Section 2.2 describes the experiment setup and analysis procedures. Section 2.3 presents results, and Section 2.4 concludes the chapter with a discussion.

■ 2.1 Basic Definitions

Previously we had described a phoneme as the sound of a consonant or vowel, and phoneme rate was the rate at which a person spoke a particular consonant or vowel. In this section, we formalize our terminology.

A phone is a speech sound, and it is the shortest speech sound distinguishable from other speech sounds based purely on the acoustic properties of the sound [132]. A set of phones has been defined by the International Phonetics Organization Association (IPA) to represent essentially all the sounds found among all the world's languages. Each language only uses a subset of phones from among all possible phones [6].

A phoneme is different than a phone. A phoneme is the shortest speech sound that distinguishes one word from another within a language. Whereas a phone is constant across languages, a phoneme is language dependent. One or more phones may correspond to the same phoneme [28].

This thesis recognizes 39 phonemes for the English language and a 40^{th} phoneme to represent silence. The IPA has designated symbols for each phone in its alphabet, but for computational ease, this thesis represents phonemes using the ASCII friendly Arpabet notation of one or two alphabetic characters per phoneme. An example of each of the phonemes along with a word that uses the phoneme in Arpabet notation is in

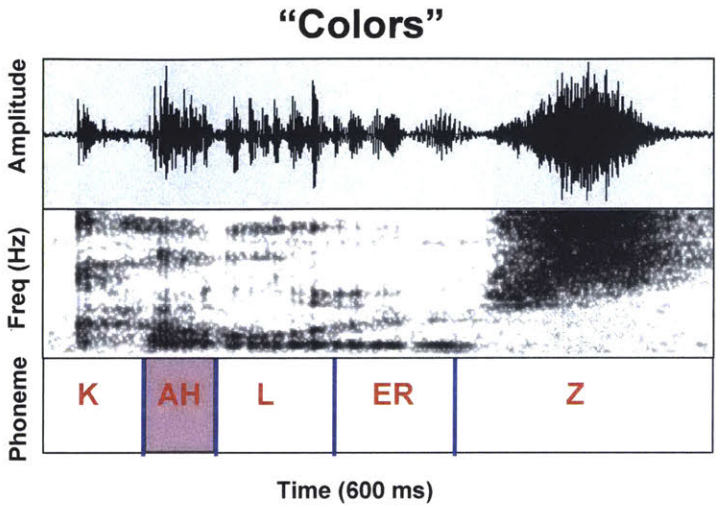


Figure 2.1: Phoneme annotated example of the word "colors." The top row shows the amplitude vs time of the utterance. The middle row shows the spectrogram of frequency vs time with darker regions corresponding to greater energy. The bottom row shows each automatically identified phoneme label and its phoneme boundaries in Arpabet notation.

Table 2.1 from the Carnegie Mellon University (CMU) pronunciation dictionary [81]. Arpabet phonemes will be represented with capital letters, and they will be surrounded by square brackets outside of tables and figures. We represent the "silence" phoneme as [H#] or with an asterisk, [*].

Phonemes are identified automatically from spoken utterances using a proprietary phoneme recognizer [118]. An example of the output of the program is shown in Figure 2.1.

The phoneme recognizer solves two challenges. It must identify the phoneme start and stop boundaries and label the phoneme itself. This is in contrast to other forms of phoneme alignment or recognizers that use a transcript of the text to know which phonemes to identify in which order. These other forced alignment algorithms then only need to solve the boundary placement problem. We chose to use a transcript-free phoneme aligner because in practical cases a transcript is not going to be available. Furthermore, a transcript assumes that all individuals pronounce a word and all its phonemes according to the dictionary's pronunciation. Regional dialects and speaking habits can cause the same word to be pronounced with different phonemes. A forced alignment program can mistakenly attempt to identify boundaries of phonemes that are not present as well as miss inserting phonemes that were not in the forced aligner's transcript.

Table 2.1: Examples from the CMU pronunciation dictionary of the 39 phonemes and silence (the 40^{th} "phoneme") in this thesis [81].

Arpabet Phoneme Abbreviation	Example Word	Example Pronunciation
AA	odd	AA D
AE	at	AE T
AH	hut	HH AH T
AO	ought	AO T
AW	cow	K AW
AY	hide	HH AY D
B	be	B IY
CH	cheese	CH IY Z
D	dee	D IY
DH	thee	DH IY
EH	Ed	EH D
ER	hurt	HH ER T
EY	ate	EY T
F	fee	F IY
G	green	G R IY N
HH	he	HH IY
IH	it	IH T
IY	eat	IY T
JH	gee	JH IY
K	key	K IY
L	lee	L IY
M	me	M IY
N	knee	N IY
NG	ping	P IH NG
OW	oat	OW T
OY	toy	T OY
P	[pea]	P IY
R	read	R IY D
S	sea	S IY
SH	she	SH IY
T	tea	T IY
TH	theta	TH EY T AH
UH	hood	HH UH D
UW	two	T UW
V	vee	V IY
W	we	W IY
Y	yield	Y IY L D
Z	zee	Z IY
ZH	seizure	S IY ZH ER
H#, *	<silence>	

Phoneme rate is the rate at which an individual phoneme is spoken. To compute the phoneme rate for the j^{th} phoneme in our alphabet of 40 phonemes, we count the total number of times the phoneme is spoken over the course of an utterance and divide by the total amount of time spent saying that phoneme. Let D be the set of durations of every instance of every phoneme in the utterance, let D_j be the subset of D that is the set of durations of every instance of phoneme j, and let d_{ij} be the duration of the i^{th} instance of the j^{th} phoneme. Then the phoneme rate for phoneme j is

$$r_j = \frac{|D_j|}{\sum_i d_{ij}} \tag{2.1}$$

where $|\cdot|$ represents cardinality of the set.

The reciprocal of phoneme rate is mean phoneme duration, which is the average amount of time spent pronouncing each phoneme. Because phoneme duration is measured per phoneme, taking the mean is just one summary descriptor for the collection of durations associated with a given phoneme. Other summary statistics include the median, standard deviation, interquartile range, maximum, and minimum.

In addition to phoneme rate, there are two other important measures of speech rate: speaking rate and articulation rate. Speaking rate is the total number of non-silent phonemes spoken divided by the total time of the entire utterance, including the silence between words and sentences:

$$r_{speak} = \frac{\sum_{j=1}^{39} |D_j|}{\sum_{j=1}^{40} \sum_i d_{ij}} \tag{2.2}$$

By contrast, articulation rate is the same as speech rate except total duration excludes silence:

$$r_{art} = \frac{\sum_{j=1}^{39} |D_j|}{\sum_{j=1}^{39} \sum_i d_{ij}}. \tag{2.3}$$

■ 2.2 Methods

In this section, we cover our experimental protocol and analysis procedures for characterizing phoneme rate as a biomarker of depression.

■ 2.2.1 Protocol

Subjects read aloud "My Grandfather" [35, 138], "The Rainbow" [42], and "The Caterpillar" [99] into a laptop's built-in microphone at a self-paced speed and volume. These three passages were selected for their wide use and long history in the speech community. In particular, they have been used for tracking neuropsychological disorders through the voice. Each subject also completed a Beck Depression Inventory II questionnaire to determine depression severity. A BDI score greater than or equal to 14 was used

Table 2.2: Linguistic characteristics of the read passages determined from readabilityscore.com.

Attribute	Rainbow	Caterpillar	Grandfather
Nominal Speaking Time (mm:ss)	2:38	1:43	1:02
Sentiment	Neutral (Slightly Positive)	Neutral (Slightly Positive)	Neutral (Slightly Positive)
Flesch-Kincaid Grade Level	7.8	4.8	6.4
Sentence Count	19	16	8
Words per Sentence	17.4	12.2	16.4
Syllables per Word	1.4	1.3	1.3

as the depression cutoff [10]. Additionally, each subject completed a battery of other questionnaires that were not analyzed as part of this study. Study data were collected and managed using REDCap electronic data capture tools hosted at MIT [64, 97], and this study had MIT Institutional Review Board approval.

Our aims with the read passages were to identify which of the standard read passages is best used to identify depression and to understand why a passage is successful for identifying depression. These passages may differ in emotionality, length, and complexity of writing. A highly emotional passage may cause different responses from depressed or control subjects. Longer passages provide more speech over which to compute speech features, so we would anticipate estimates would be less variable. Complex writing both in terms of word and sentence structure will have more demanding motor requirements than simple words and sentences. Therefore, the more emotional, longer, and more complex a passage, the greater the accuracy with which we would expect to classify depressed vs control subjects.

We characterize these attributes using an available analysis service, readabilityscore.com. The results from this service for the Grandfather and Rainbow passages agreed well with a published study by Ben-David et al. [11]. Unfortunately, the Caterpillar passage was not part of Ben-David et al. [11]'s study, so we present in Table 2.2 all the results from readabilityscore.com for comparability across all our passages.

We observe that all three passages have similar syllables per word at around 1.3, and similar sentiment, which is essentially neutral. However, they differ markedly in terms of their nominal speaking time and in their grade level difficulty. The Rainbow passage

is more than twice as long as the Grandfather passage and about fifty percent longer than the Caterpillar passage. Because of their similar emotional content but varying lengths we can test for the influence of passage duration on depression classification performance.

We used the Flesch-Kincaid (F-K) system to evaluate reading difficulty [74]. The F-K system is a weighted sum of words per sentence and syllables per word, and it is widely used for evaluating the difficulty of comprehension of written material. We observe that the Rainbow passage is at a 7.8 US grade level vs the 6.4 grade level of the Grandfather passage and the 4.8 grade level of the Caterpillar passage. The small disparity in reading levels may be a confound to passage length, but given that all our subjects are adults, we do not anticipate this to be an influential variable.

■ 2.2.2 Vocal Biomarkers

Phonemes were automatically identified and segmented using an automatic phoneme recognizer described previously. For each of the 39 phonemes and silence, which we counted as a 40^{th} phoneme, we computed the mean phoneme duration. We also computed a phoneme based speaking rate and articulation rate.

■ 2.2.3 Direct Comparisons with Prior Art

We have mentioned that phoneme rate was first explored by Trevino et al. [129], and then later by Williamson et al. [144]. In this section, we follow their respective methodologies to facilitate direct comparison of our study's results with their findings.

Comparison with Trevino et al. [129]

Trevino et al. [129] studied 35 subjects in an English free speech task whose depression severity was quantified by a clinician administered Hamilton Depression Scale assessment (HAMD). This study's metric differed from our use of the Beck Depression Inventory-II (BDI), which is a self-report score. We chose to use the BDI over the HAMD because self-report scores can scale to assessing patients worldwide since a clinician is not required. Our phoneme recognizer is the same recognizer employed by Trevino et al. [129] in their study.

Trevino et al. [129] created an aggregate phoneme rate biomarker based on correlations of individual phoneme rates with the HAMD. For each phoneme, the phoneme's mean duration was computed per subject. Then, these subject specific mean durations were correlated per phoneme against the subject HAMD scores using a Spearman correlation. If a phoneme was significantly correlated ($p < 0.05$) with the total score, it was added to the set of phonemes that would be aggregated. The sign of the correlation was also noted. Once all the phonemes to be aggregated were identified, the phonemes were combined per subject by creating a signed weighted sum of the mean phoneme durations. Weights were either $+1$ or -1 depending on the sign of the phoneme's univariate Spearman correlation. The formula for the aggregate or fused mean phoneme

duration biomarker, $L_{Trevino}$, is

$$L_{Trevino} = \sum_{n=1}^{n_{sig}} \alpha_n L_n \tag{2.4}$$

where n_{sig} is the number of significantly correlating phonemes, α_n is the sign of the Spearman correlation, and L_n is the mean duration of the n^{th} significantly correlating phoneme.

Both univariate significance testing and aggregation were also computed by Trevino et al. [129] to determine an aggregate mean phoneme duration for the psychomotor retardation (PMRT) HAMD subscore. Our study did not have the HAMD available, so we followed the same methodology but correlated against the BDI.

Comparison with Williamson et al. [144]

Williamson et al. [144] studied phoneme rate among other vocal biomarkers of depression in a population of German speakers. In the analyzed speech, the subjects read one German passage in German, and also responded in German to a free response set of questions. The depression severity metric was the Beck Depression Inventory II, and Williamson et al. [144] used the same phoneme recognizer as Trevino et al. [129] and our study. Williamson et al. [144] followed a similar but not identical procedure to Trevino et al. [129] in deriving an aggregate phoneme duration measure. Rather than performing a significance test, they picked the top phonemes with the largest absolute values of a univariate Pearson correlation with the BDI. They then combined these phoneme durations using a signed weighted sum. They used the sign of the Pearson correlation and the magnitude of the correlation when computing the weight. Their formula is

$$L_{Williamson} = \sum_{n=1}^{n_{top}} \frac{\text{sign}\,(R_n)}{1 - R_n^2} L_n. \tag{2.5}$$

n_{top} was chosen as 6 for the read passage and 10 for the free speech passage. We show results using the top 6 phonemes.

■ 2.2.4 Machine Learning

To evaluate the utility of phoneme rate as a biomarker of depression, we must move beyond the correlation analyses just discussed. In those analyses, we were only assessing whether the phoneme rate biomarker correlates with depression severity. However, those correlations provide an optimistic view of the biomarker's utility in practice. To assess how well the biomarker would do in classifying a subject as depressed or not depressed, we must construct a classifier that returns a binary decision on a previously unseen subject.

Ideally, we would have so many subjects that we could construct a classifier on a training set of subjects and then test the classifier's performance on a held out test set

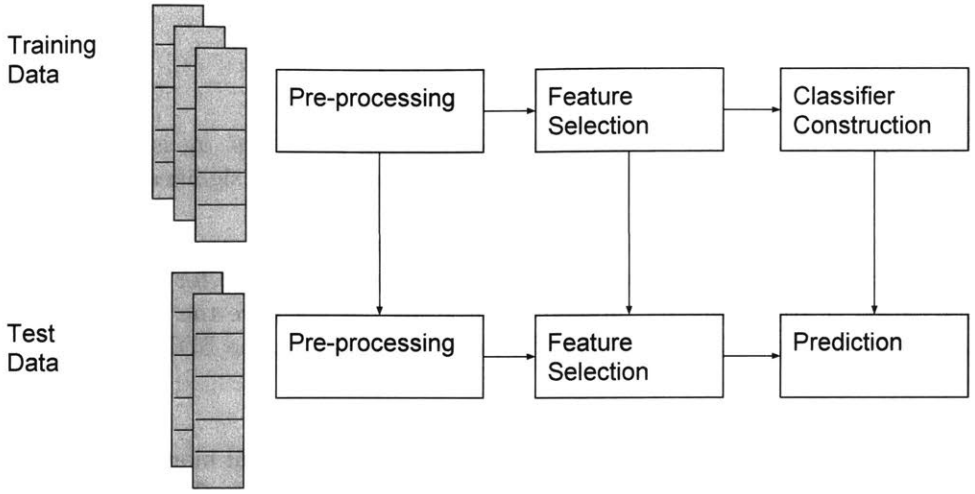


Figure 2.2: The machine learning pipeline operates within a shuffle-split cross-validation loop, so parameters from the processing stages that operate on training data (top row) are applied to the test subjects (bottom row). Raw feature vectors are pre-processed before feature selection, feature fusion, and classifier construction.

of subjects. Instead, our small sample size drives us to use shuffle-split cross-validation. Shuffle-split cross-validation splits the available data into a train set and test set. The classifier is trained on the training set and tested on the test set to derive performance metrics. Then, the entire dataset is randomly split again into a train set and test set. The new train set will partially overlap with the previous train set, and the new test set will partially overlap with the previous test set. Shuffle-split differs from K-fold cross-validation in that the test sets in shuffle split have overlapping subjects. However, shuffle split allows many more folds of the data to be tested by relaxing this constraint, and consequently provides a more complete picture of the variability of performance than K-fold cross-validation.

Our machine learning pipeline consists of the following stages: pre-processing, feature selection, classifier construction, and classifier performance evaluation. Each of these steps is performed within a shuffle-split cross-validation loop, so, for example, pre-processing parameters are derived using all the training subjects, and then the pre-processing parameters are applied to the test subjects. Figure 2.2 shows a graphical description of the process that we now describe in detail.

Pre-processing through Classification

For a single shuffle-split fold, we divide the dataset such that 83 percent of the subjects are used for training and 17 percent of the subjects are used for testing (for a 30 subject

dataset, these numbers correspond to 25 subjects and 5 subjects). Furthermore, to avoid classifier training bias towards a class that may be underrepresented in the training set, the number of samples of the over-represented class in the training set is randomly reduced to equal the number of samples of the other class. For each subject, for each passage we create a 40 element feature vector consisting of the mean phoneme durations. If a phoneme was not spoken by a subject, the duration is set to 0. On the training set, we compute normalization constants such that the features have zero mean and unit variance after normalization.

We then select the most informative features by applying an extremely randomized trees [52] algorithm implemented in scikit-learn [100] to the training data. This algorithm is a form of a random forest algorithm in which each split point is chosen from a random selection of splits rather than an exhaustive search. Features that are consistently chosen across the ensemble of trees, and features that are consistently chosen at shallow depths are more important than infrequently chosen features and features at deep tree depths. All features that have an importance greater than the mean feature importance are retained and are used by the subsequent classifier for training.

The classifier is another extremely randomized trees ensemble. It is trained only on the features that were selected by the feature selection step. The final output from this classifier is a real valued vote between 0 and 1 as to whether a subject is depressed or not. A vote is a measure of how confident the classifier is with respect to its decision (0: subject is likely a control, 1: subject is likely depressed). We use 1000 trees for our feature selection model, and we use 2000 trees for the classification model. These values were empirically chosen from prior experience with this algorithm.

We use 300 shuffle-split cross-validation folds in total. We divide these splits into 10 partitions of 30 folds. Each partition contains multiple votes for the same subject because each test subject may appear in several folds given that each fold randomly selects five subjects for testing (assuming a 30 subject dataset). The mean of the votes for a subject is taken as the final score for the subject, so each subject in the partition has an associated score (a real number) and a binary class label.

Evaluation

Accuracy is the most intuitive metric for evaluating a classifier's performance. However, accuracy is not appropriate when the number of subjects in each of the two classes is unbalanced, as in the case of this study. Therefore, we use the Receiver Operating Characteristic (ROC) schematically shown in Figure 2.3.

The x-axis of a ROC curve is the false positive rate also known as (1 - specificity). The false positive rate is the probability of declaring a person depressed when the person is a control. The y-axis is the true positive rate also known as the sensitivity or the recall. The sensitivity is the probability of correctly declaring a person depressed when the individual is depressed. Any point on the ROC curve shows the trade-off between sensitivity and specificity for a given threshold.

The area under the curve (AUC) provides a summary measure of overall algorithm

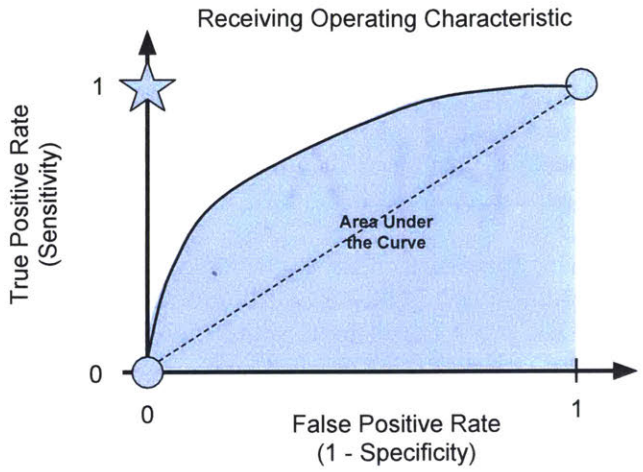


Figure 2.3: Schematic example of a Receiver Operating Characteristic (ROC) curve.

performance. If the AUC is greater than 0.5 (i.e., the majority of the curve lies to the upper left of the diagonal line running from (0,0) to (1,1)), then the algorithm is performing above a baseline of random guessing. Ideal performance would be perfect sensitivity and perfect specificity, which is represented by the star in the figure.

We compute the AUC for each of the partitions by sweeping the threshold from 0 to 1 on the subject classifier scores. We then compute statistics on the AUC values determined from all the partitions. With shuffle-split folds, not all subjects are chosen the same number of times in the test set as this is a random process.

■ 2.2.5 Stability of the Phoneme Features

We determine the most discriminative features for each of the passages based on the feature selection process just described. Additionally, we compare the selected features for each of the passages against each other. If the phoneme rate biomarker is insensitive to the read passage, we can infer that the phoneme rate biomarker is tapping into a fundamental motor characteristic associated with certain phonemes. However, if the phoneme rate biomarkers are different among the passages, then it may be that the passages are too short to highlight particular phonemes best suited for depression. Instead, the word context of the phonemes has a greater influence on the discriminative value of the phoneme than the phoneme itself. For example, co-articulation effects may alter how an [AH] is pronounced in a word that appears in the Caterpillar passage but not the Grandfather passage. Consequently, within the context of that word, [AH] may be a biomarker of depression whereas within the Grandfather passage, [AH] may not be a useful feature.

We evaluate feature stability by identifying the most frequently selected N phonemes for each of the three passages. We also report the importance of each of the phonemes averaged across all the shuffle-split folds. If a feature was not selected for a particular fold, it was assigned an importance of 0 for that fold.

■ 2.3 Results

This section parallels the organization of the Methods in Section 2.2. We present dataset characteristics (2.3.1), compare our results to prior studies (2.3.2), present classification performance (2.3.3), and present the feature stability analysis (2.3.4).

■ 2.3.1 Dataset Characteristics

32 subjects completed the Grandfather passage, Rainbow passage, and Caterpillar passages. However, subject 961's Caterpillar passage was not saved and subject 880 read an abridged version of the Rainbow passage. To facilitate comparisons among the three passages, these subjects were excluded from analysis.

To assess accuracy of the phoneme recognizer, we created Praat textgrid files for the phonemes and plotted the phonemes against the spectrograms in approximately 2 second intervals. The spectrograms and aligned phonemes were then visually inspected for egregious, consistent failure to identify speech segments vs non-speech segments. We also checked for approximately correct boundaries by checking for correspondence between phoneme boundaries and changes in the spectrogram. We checked the first and last few words/phonemes for each of the passages to obtain a general sense of the accuracy of the phonemes and not just the accuracy of the boundaries of the phonemes.

Some notable points of difficulty for the phoneme recognizer were the first and last few phonemes of passages because sometimes the read passages were parsed in such a way as to have no or too little silence before speech onset. The segmentation of the passage from the audio was performed by one analyst for the Rainbow and Caterpillar passages and another analyst for the Grandfather passage, and they had different styles for onset/offset marking for the passages. Another point of difficulty was identifying the "Tick, tick, tick" segment of the Caterpillar as speech for some subjects. While there was a range of accuracy among subjects as qualitatively assessed, we did not exclude any subjects as having too few phonemes to perform analysis.

Demographics for the analyzed subjects are presented in Table 2.3 and Table 2.4. The interquartile range, IQR, is the difference between the 75^{th} and 25^{th} percentile, and the median is the 50^{th} percentile of the data[1].

We show as a plot the individual variability of the length of time each subject took to complete the read passages in Figure 2.4[2]. The Rainbow passage takes the most time to complete, followed by the Capterpillar passage, and then the Grandfather passage. This ordering holds for all subjects.

[1] cobidas_prescan.ipynb
[2] read_dur_plot.ipynb

Table 2.3: Number of subjects (Count) in the MIT read passage dataset, and the Beck Depression Inventory statistics for the subjects. IQR: Interquartile range.

		Count	Mean	Std. Dev.	Median	IQR	Min.	Max.
	Sex							
Control	F	3	0.33	0.58	0	0.5	0	1
	M	7	1.71	1.70	1	2.5	0	4
Depressed	F	11	25.64	10.01	26	15.0	14	44
	M	9	30.78	6.48	31	11.0	21	39

Table 2.4: Number of subjects (Count) in the MIT read passage dataset, and the age (years) statistics for the subjects. IQR: Interquartile range.

		Count	Mean	Std. Dev.	Median	IQR	Min.	Max.
	Sex							
Control	F	3	24.67	4.93	27	4.5	19	28
	M	7	22.57	1.40	23	1.5	20	24
Depressed	F	11	29.18	11.97	27	15.5	18	52
	M	9	33.11	12.64	27	22.0	18	51

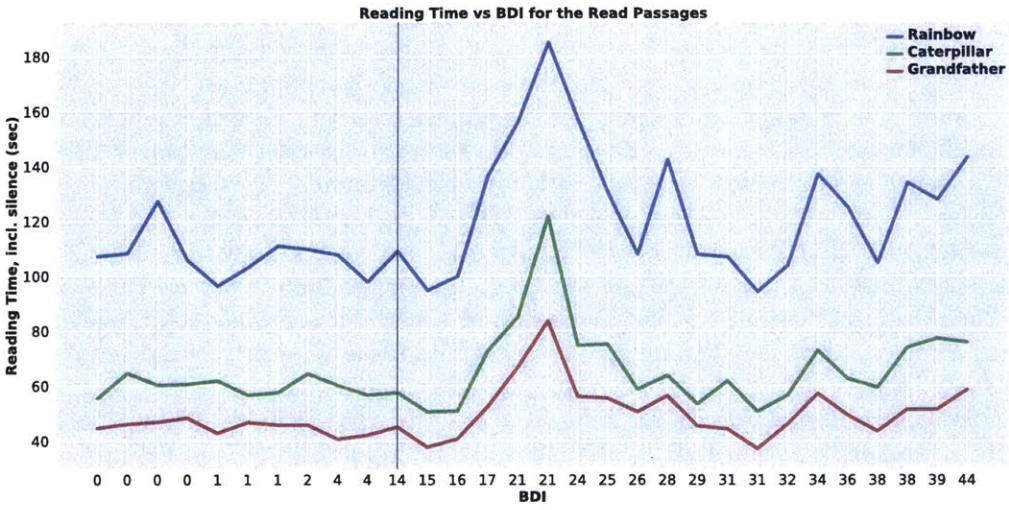
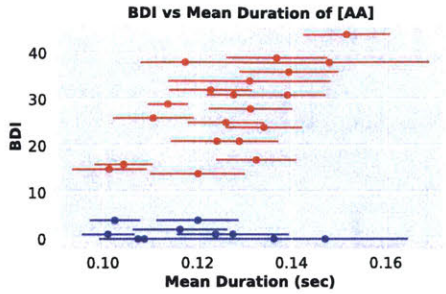
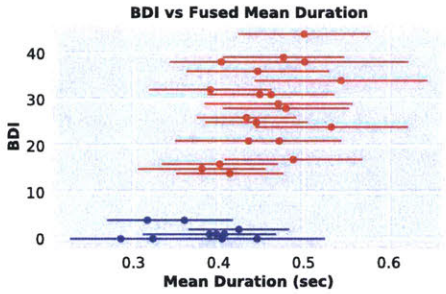


Figure 2.4: Reading time vs BDI for the different passages. The Rainbow consistently takes the longest time to read, and the Grandfather takes the least amount of time. The x-axis shows one tick mark per subject with the subject's corresponding BDI score. A subject with a BDI score greater than or equal to 14 is classified as depressed (vertical, gray line).



(a) Scatter plot of subject BDI scores vs the [AA] phoneme's mean duration for the Rainbow passage. The Spearman correlation is 0.44 (p=0.015).



(b) Scatter plot of subject BDI scores vs the fused mean phoneme duration for the Rainbow passage. The Spearman correlation is 0.62 (p=0.0002).

Figure 2.5: Spearman correlation of an individual phoneme's duration with the BDI (left), and the Spearman correlation with the aggregated mean phoneme duration using all individually significantly correlating phonemes after the methodology of Trevino et. al. Red indicates depressed subjects, and blue indicates control subjects.

■ 2.3.2 Comparisons with Prior Art

We show example figures of a representative individual phoneme duration and the fused phoneme rate biomarker in Figures 2.5a and 2.5b using the Trevino et al. [129] method. These scatter plots reveal that a noticeable monotonic relationship exists across the range of depression scores for a single phoneme. When the phoneme durations are fused by taking the individually significantly correlating phonemes, the overall correlation between the BDI and the speech feature becomes even stronger as expected[3]. The error bars in Figure 2.5a are plus and minus one standard error of the mean, and the error bars in Figure 2.5b are plus and minus the square root of the sum of the variances of durations of the fused phonemes.

The correlations of fused mean phoneme durations with depression severity, according to the methodologies in Trevino et al. [129] and Williamson et al. [144], are in Table 2.5 and Table 2.6 respectively[4]. We see that overall correlation performance is comparable to both prior studies. Our correlations are higher than the HAMD Total correlation for all three read passages and and higher than the correlations reported by Williamson et al. [144].

■ 2.3.3 Classification Performance

Classification performance using the phoneme rate biomarker is reported in Table 2.7. We see that performance on the Rainbow passage is the best with a mean AUC of 0.73,

[3] phn_dur_table.ipynb
[4] trevino2011_replicate.ipynb

Table 2.5: Trevino et al. [129] reported Spearman correlations of fused mean phoneme durations with the Hamilton depression total score (HAMD), and the psychomotor retardation (PMRT) subscore using the Mundt et al. [93] dataset. The table also shows Spearman correlations using the Trevino et al. [129] method on our MIT dataset. The individual phonemes that were fused are also listed. All p values are uncorrected.

Research Dataset	Speech Protocol	Depression Scale	Spearman ρ (p value)	Sig. phonemes (p <0.05)
Mundt et al. [93]	Free speech	HAMD Total	0.35 (1.8e-4) [129]	H#, S, K, IH, AA [129]
Mundt et al. [93]	Free speech	HAMD PMRT	0.58 (1.7e-11) [129]	H#, AE, IY, AY, EY, AO, OW, EH, AW, UH, ER, G, K, NG, R, S, T, V, W, Z [129]
MIT	Rainbow	BDI	0.62 (0.0002)	AH, ZH, SH, L
MIT	Caterpillar	BDI	0.49 (0.007)	AE, P
MIT	Grandfather	BDI	0.48 (0.008)	M

Table 2.6: Williamson et al. [144] reported Pearson correlations of fused mean phoneme durations with the BDI scale on the Audio Visual Emotion Challenge (AVEC) depression dataset [133]. The table also shows Pearson correlations using the Williamson et al. [144] method on our MIT dataset. The top phonemes reported by Williamson et al. [144] are listed, and the top 6 phonemes fused for the MIT dataset are listed for each read passage. All p values are uncorrected.

Research Dataset	Speech Protocol	Depression Scale	Pearson ρ (p value)	Top 6 phonemes
AVEC [133]	Free speech	BDI	0.57 [144]	NG, T, HH, EY, OW, ER [144]
AVEC [133]	Northwind	BDI	0.54 [144]	L, AH, N, IH, B, OW [144]
MIT	Rainbow	BDI	0.64 (0.0002)	SH, ZH, AA, Y, L, B
MIT	Caterpillar	BDI	0.61 (0.0003)	P, Y, AE, B, T, AY
MIT	Grandfather	BDI	0.59 (0.001)	M, R, ER, V, JH, AO

Table 2.7: Receiver operating characteristic area under the curve (ROC AUC) depressed vs control classification performance on each of the read passages from the MIT dataset. Features were mean phoneme durations, and statistics were computed over 300 shuffle-split cross-validation folds with 30 folds per partition to create 10 partitions. Statistics are reported on the ROC AUC across the ten partitions.

Speech	ROC AUC			
Protocol	Mean	Std. Dev.	Median	IQR
Rainbow	0.73	0.03	0.74	0.04
Caterpillar	0.52	0.04	0.52	0.07
Grandfather	0.45	0.04	0.45	0.03

followed by the Caterpillar at 0.52, and then the Grandfather at worse than chance with an AUC of 0.45. Among all three passages there is minimal and approximately equal variability in performance for any given partition. The standard deviation is 0.03, 0.04, and 0.04 respectively[5]

2.3.4 Feature Stability

We report the top 10 most commonly selected phonemes in Table 2.8 along with each phoneme's importance[6]. Recall that the Rainbow passage performed the best overall in terms of classification. Among the top ten most frequently selected phonemes, there is partial overlap among the passages. The Rainbow and Caterpillar passage share [F], [IH], [T], the Rainbow and Grandfather passages share [IH], [B], [T], [V], the Caterpillar and Grandfather passages share [OW], [IH], [M], [T], and the intersection of all three passages contains [IH], and [T].

2.4 Discussion

We discuss our results in the context of this chapter's opening research questions. The most important take away from this chapter should be the viability of the phoneme rate biomarker for assessing depression. In cross-validation experiments, we do see that the phoneme rate biomarker is able to predict depression severity with an AUC score of 0.73. However, this result is conditioned on having sufficient speech over which to estimate the phoneme rate biomarkers for model training and testing. In answer to the first question in this chapter, how does performance change with amount of speech, we might say that longer speech samples result in better performance. We recommend that a minimum of 100 seconds should be collected per subject. This minimum is based on the approximate duration of the Rainbow passage which gave above chance results.

However, we pointed out that the passages due differ in terms of complexity. We

[5] pipe_sg_np.ipynb, pipe_fusion.ipynb

[6] pipe_feat_impt.ipynb

Table 2.8: Top 10 most frequently selected phonemes and each phoneme's importance. Phonemes are listed from most frequently selected (top) to least frequently selected (bottom). Importances are normalized such that the most important phoneme across all phonemes and all passages has an importance of 1.0.

Rainbow			Caterpillar			Grandfather		
Phoneme	Freq.	Impt.	Phoneme	Freq.	Impt.	Phoneme	Freq.	Impt.
ZH	0.99	1.00	T	0.95	0.72	V	0.96	0.90
V	0.83	0.68	G	0.85	0.64	R	0.82	0.69
L	0.78	0.49	AE	0.83	0.67	B	0.79	0.58
F	0.72	0.41	Y	0.80	0.59	T	0.74	0.56
B	0.71	0.51	F	0.78	0.55	HH	0.68	0.51
IH	0.70	0.46	W	0.63	0.38	M	0.67	0.47
SH	0.69	0.47	M	0.56	0.34	ER	0.60	0.37
CH	0.66	0.43	IH	0.55	0.33	IH	0.48	0.30
AW	0.60	0.32	OW	0.55	0.34	JH	0.46	0.29
T	0.57	0.37	EH	0.53	0.32	OW	0.44	0.26

reduced the duration of the Rainbow passage to half for each subject, keeping the first half of all the phonemes in one dataset and keeping the second half of all the phonemes in another dataset. The mean and standard deviation ROC AUC for these "halved" sets is 0.82 (0.02) and 0.74 (0.04) respectively. Consequently, complexity of the Rainbow passage may actually be the key attribute of the passage rather than its length.

Our second research aim was to determine whether or not particular phonemes were biomarkers that were robust to passage content. Ideally, distinguishing features should not depend on the passage because that would indicate certain phonemes truly are reflecting neuromotor disturbances associated with depression. Unfortunately, we see that the top ten most frequently selected phonemes among the passages only partially overlap, and all three passages only share two common phonemes.

Furthermore, when we compare the most frequently selected phonemes as derived by classification performance compared to the phonemes selected by univariate regression as done by Trevino et al. [129] and Williamson et al. [144], we see some overlap. The Rainbow passage has [SH], [ZH], and [L] selected by all three approaches, the Caterpillar passage has [AE] selected by all approaches, and the Grandfather passage has [M] selected by all approaches.

Consequently, we must conclude that we do not have sufficient amounts of speech per subject to assess passage independence for phoneme biomarkers. Instead, passage dependent effects are contributing to classification performance. The passage dependence of the phoneme biomarker has significant implications for read speech studies as it strongly suggests that a single standard passage should be used.

Several methodological considerations should be kept in mind when interpreting the

correlations in Tables 2.5 and 2.6 in which we compared our correlating phonemes to those in prior work. First, Trevino et al. [129] data was free speech not read speech. Second, the phoneme recognizer used in Williamson et al. [144] was trained in English but applied to German read and free speech. It may be that the phonemes identified are different than if a German training corpus had been used for the phoneme recognizer. Third and most important when assessing overall performance, the correlations are best fit correlations. The results reported in the table are not cross-validated. They are derived from a simple Spearman or Pearson correlation of the aggregate phoneme biomarker with the depression score. We reported our results in the same way for comparison, but it should be noted that this method overestimates the predictive power of the aggregate feature since effectively this procedure is akin to training and testing on the same dataset without cross-validation.

For context on the variability in the scatter plots, we included bars that showed extent of variability of the phoneme durations. While we do not have a direct measure of variability in reported BDI scores, we report from the literature that the BDI-II has a one week test-retest correlation of 0.93 ($p < 0.001$) [10] in a 26 subject outpatient sample. By contrast, in a non-clinical population, Ahava and Iannone [2] found a forty percent decrease in scores over an eight week period in a cohort of 150 subjects. However, Ahava and Iannone [2] used the BDI-IA rather than the BDI-II form of the survey.

In this chapter, we have contributed a thorough analysis of the phoneme rate biomarker in read speech and seen that it holds promise for assessing depression. We replicated correlations observed in prior research to establish the presence of the phoneme rate biomarker. We also showed that the discriminating phonemes across passages are variable, and that this should be kept in mind when creating a depression assessment protocol. In the following chapter, we will probe the neural basis for this biomarker with the aim of using an understanding of it to better assess depression.

Chapter 3

Task-Based fMRI Analysis of Phoneme Rate

THIS chapter describes a functional magnetic resonance imaging (fMRI) investigation of the neural correlate of phoneme rate in depressed speech. Using overt production of read, emotional sentences by a depressed and control population, we answer two questions. Do emotional sentences influence phoneme rate production in the brain? Is there a difference in how this influence occurs between depressed and control subjects? Our research hypothesis is that depression interacts with phoneme rate in the basal ganglia.

Section 3.1 provides background material on speech neuroanatomy, depression, and the technical challenges of using fMRI for speech production studies. It also motivates our research hypothesis and discusses the novel aspects of this study. Section 3.2 covers the experimental methods, general linear model (GLM) results, and discussion. We then turn to a connectivity analysis using dynamic causal modeling (DCM) in Section 3.3 as an alternative means of testing our hypothesis.

■ 3.1 Background and Research Hypothesis

We open with a basic primer to speech neuroanatomy (3.1.1) and current neurobiological models of depression (3.1.2). The background material provides context for a discussion of the main research hypothesis (3.1.3). Section 3.1.4 highlights novel aspects of the study. Section 3.1.5 discusses the particular strengths and challenges of using fMRI in overt speech production studies which will aid in understanding our imaging protocol and interpreting our imaging results.

■ 3.1.1 Neuroanatomy of Speech Production

Speech production results from the coordinated activity of a distributed set of neuroanatomical regions. These regions include cortical and subcortical components as well as the cerebellum and brain stem and can be seen in Figure 3.1. The cortex is heavily folded, and the outer portion of the fold is called a gyrus, while the parts of the cortex that have been folded inward and are not visible are called sulci (plural of

sulcus). The main cortical landmarks pertinent to this thesis include several prominent gyri (plural of gyrus) including the inferior frontal gyrus (IFG), the pre and post central gyrus, and the superior temporal gyrus (STG). The top of Figure 3.1 shows these gyri as seen from the lateral, outside view of the brain [13].

If the brain is cut into two hemispheres, one can look at the medial side of a hemisphere instead of the lateral side as in the middle of Figure 3.1. The medial view shows the supplementary motor area (SMA) and the cingulate gyrus [13].

Within a hemisphere, we are interested in several buried structures denoted by dotted lines in the bottom of Figure 3.1. The basal ganglia is the name for a group of structures that includes the putamen, caudate, and nucleus accumbens. The putamen and caudate can be grouped together because of their cellular structure and how they develop embryologically. The caudate and putamen together are called the striatum. At the anterior portion of the putamen, where the putamen and caudate meet, there is a structure called the nucleus accumbens. The nucleus accumbens and the overlap of the caudate and putamen with the nucleus accumbens are termed the ventral striatum [13].

Research into speech production has assigned approximate functional roles to these different regions. The inferior frontal gyrus can be further divided into the pars opercularis, pars triangularis, and pars orbitalis segments moving posterior to anterior along the IFG. Pars opercularis (IFo) and pars triangularis (IFt) correspond to a region approximately termed Broca's area. IFo, IFt, and precentral gyrus are responsible for creating and executing articulatory gestures. The posterior superior temporal gyrus is responsible for auditory self-monitoring [60].

The basal ganglia is an action selection mechanism which means the basal ganglia decides which motor program to execute and triggers the cortex and cerebellum to perform the execution by way of the thalamus. The basal ganglia in conjunction with the SMA is hypothesized to play a role in the sequencing of speech chunks. A speech chunk is an over-learned speech motor program that could be as elemental as a phoneme or as complex as a short multi-syllabic utterance. We use "over-learned" as an adjective because the basal ganglia is hypothesized to be more involved in practiced motor commands and less involved in novel motor commands [60].

The basal ganglia, specifically the caudate, and SMA orchestrate the chaining together of these chunks. The putamen participates in executing those chunks [16]. The thalamus acts as the relay center between cortical and subcortical processing via cortical-basal ganglia-thalamus-cortical loops. It is involved in chunk planning and chunk execution. For detailed reviews of speech production, we refer the reader to Guenther [60], Price [108], Indefrey and Levelt [71], and Hickok and Poeppel [66]. For an excellent introduction to neuroanatomy and function in general, we recommend Blumenfeld [13].

As we shall see in the next section, some of these speech regions overlap or are connected with neuroanatomical regions associated with depression (e.g., the amygdala). These connections may provide an opportunity for non-speech processes to influence

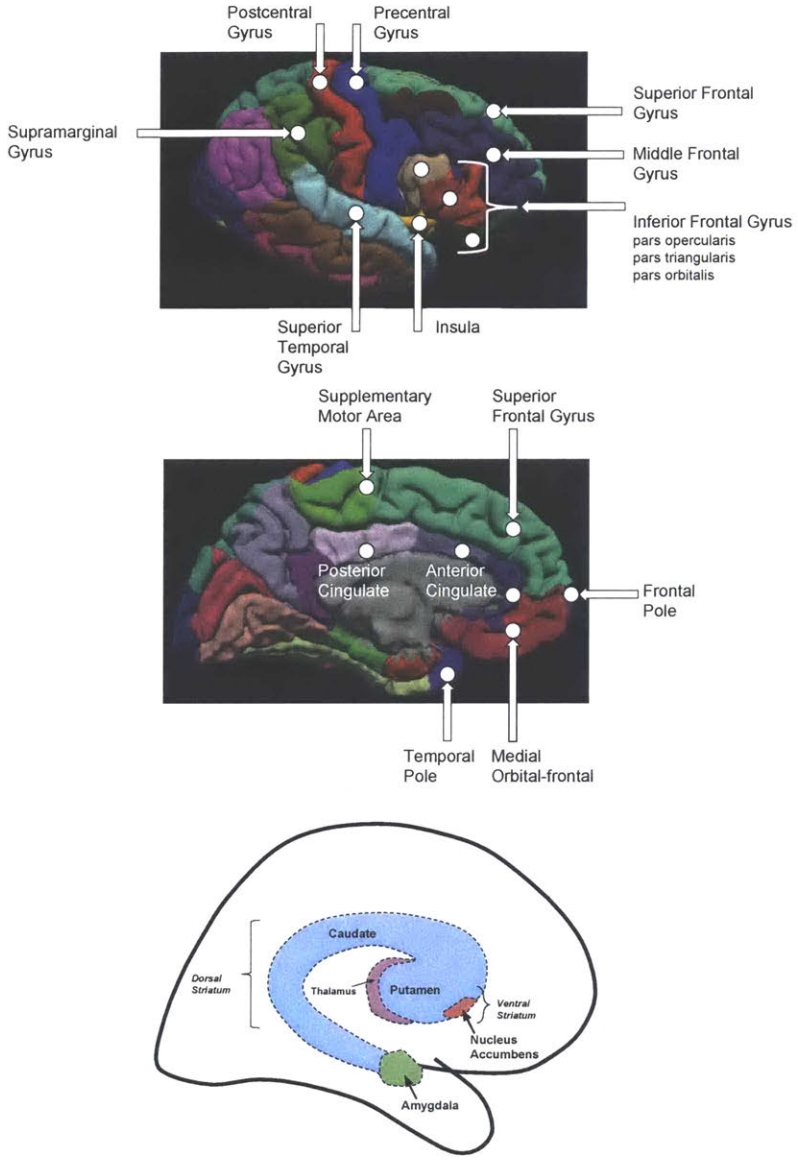


Figure 3.1: Brain regions associated with speech and depression. (Top) Selected cortical regions of interest shown on the lateral (outside) surface of the brain. (Middle) Selected cortical regions on the medial surface. (Bottom) Selected subcortical regions. Figures after [13, 44, 60].

speech production. The hypothesized modulation of the speech network by non-speech processes provides the guiding motivation for why speech can be a biomarker of neuropsychological disorders in general and depression in particular.

3.1.2 Limbic System Dysfunction and Depression

Depression can be characterized as a dysfunction of the limbic system according to Mayberg [87]. The limbic system is a set of cortical and subcortical structures that process emotional stimuli and participates in responses to them by stimulating other brain structures such as the hypothalamus to release horomones [13]. The limbic system is comprised of the hippocampus, amygdala, and several supporting elements including the cingulate cortex.

The first key idea in this thesis is that activity in the limbic system can influence motor actions. The amygdala is a processing element of the limbic system, and it has many bidirectional connections to other limbic and non-limbic components. One route by which the limbic system, through the amygdala, can influence motor actions is the ventral amygdalofugal pathway (VAP). The VAP connects the amygdala to the prefrontal cortex, orbitofrontal cortex, and anterior cingulate cortex. Additional VAP connections from the amygdala lead to the nucleus accumbens. Recall that the nucelus accumbens is part of the basal ganglia and the basal ganglia is part of speech motor control. Because the nucleus accumbens shares a boundary with the putamen and the caudate, we hypothesize that activity in the nucelus accumbens can influence activity in the caudate and putamen. Therefore the VAP is a link between the limbic system and modulation of motor actions [147].

The second key idea is that the amygdala functions differently in depressed individuals and controls. Increased amygdala activity, either laterally or bilaterally has robustly been identified as a differentiator between depressed and healthy control subjects using an emotional face matching paradigm or other emotionally valenced stimuli while undergoing functional magnetic resonance imaging (fMRI) [62, 63]. Therefore, depression influences the amygdala, and the amygdala influences motor actions, so depression can influence motor actions.

The sensitization of the amygdala to emotional stimuli due to depression motivates the use of implicitly emotional sentences in our fMRI protocol. We reasoned that emotional sentences would accentuate the limbic processing differences between depressed and control subjects and by extension accentuate the measurable effect of the limbic system's modulation of the speech network.

3.1.3 Research Hypothesis

To review, a neuroanatomical pathway exists between the limbic system, which we have seen is associated with depression, and the speech network. Specifically, this link is between the amygdala and the basal ganglia. This connection motivates our hypothesis that phoneme rate, a speech sequencing phenomenon, is modulated by the amygdala in the basal ganglia, specifically the ventral striatum. Our specific research hypothesis

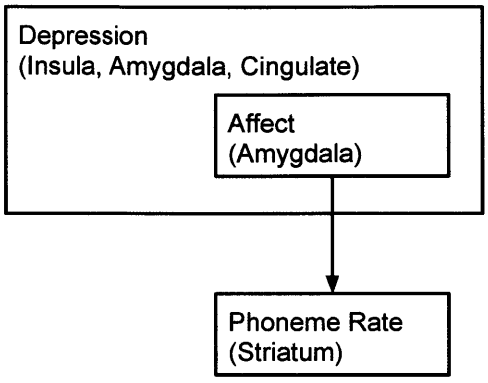


Figure 3.2: Through the amygdala's membership of a larger depression network, depression severity may modulate amygdala activity and by extension modulate the phoneme rate. Succinctly, the amygdala is hypothesized to affect phoneme rate in the caudate.

is that depression shows an interaction with magnitude of emotional sentence valence and with phoneme rate in the striatum. If we see regions of the striatum that show a depression by valence interaction overlapped with striatal regions that show a depression by phoneme rate interaction, we will have garnered support for our model in Figure 3.2.

We highlight emotional valence in our hypothesis because of the extensive prior work on valence and depression as well as related disorders like anxiety disorders, but we acknowledge that intensity or magnitude of sentence valence may be more important than sentence valence. We first analyzed our data based on valence and subsequently on magnitude of valence. We report results on our final cohort using magnitude of valence.

We focus on phoneme rate as a sequencing phenomenon instead of an articulatory effect of depression or an acoustic target effect because we see the strongest overlap between the limbic system and the motor system within the basal ganglia. The basal ganglia has a role in articulation, but as mentioned, articulation is predominantly controlled by the cortex and the cerebellum. Likewise, depression may influence the acoustic target of speech production (informally, an acoustic target is how a speech-sound sounds). However, models of speech production locate the acoustic and auditory aspects of speech to auditory cortex as opposed to the basal ganglia [60].

■ 3.1.4 Study Novelty and Relationship to Prior Art

Elements of our study have appeared in prior work, but no study has integrated these aspects as we have. Pichon and Kell [106] performed an fMRI study of explicit and *acted* emotional speech production in healthy controls. They showed that dorsal and

ventral striatum are important for preparing and executing emotional speech. We differentiate ourselves from their study in that we investigate implicit emotional speech production. Because our ultimate application is to assess neuropsychological disorders through natural speech, and implicit emotional speech is closer to natural speech than acted speech, we believed an implicit emotion paradigm would activate speech regions that were more relevant to natural speech processes than acted speech.

As alluded to in the previous section, a task with an emotional element was selected because there is evidence that emotional stimuli cause different, measurable changes in brain activity between controls and depressed subjects, particularly in the amygdala [62, 63]. While we would hypothesize that there would be measurable differences of brain activity during speech production even in neutral sentences, we believed that using emotional sentences would accentuate those differences. Our study is the first to use implicit emotional sentence production in a depressed and control population.

Several previous studies have investigated motor control of speech rate [1, 111, 141]. These studies found a positive step change in cerebellar activity around a syllable repetition rate of 3 Hz suggesting the cerebellum is recruited to move speech rate beyond 3 syllables per second, and they found a positive relationship between syllable repetition frequency and brain activity in sensorimotor, intrasylvian, and mesiofrontal cortex. However, they found a negative relationship between repetition frequency and brain activity in the putamen and pallidum. Unlike these studies, our study uses speech at a natural speech rate instead of paced speech, and our study uses complex stimuli e.g., full sentences instead of the "pa" syllable. However, there is a pacing effect at the sentence level because of the sparse imaging design that will be discussed later. Therefore, our study allows speech rate control at a more complex level of planning and execution where depression may have a greater influence on the relationship between speech rate and brain activation.

Bohland and Guenther [15] investigated speech sequencing in an fMRI study of simple and complex sequences and simple and complex syllables. A complex sequence is one in which the syllables were different instead of repeated (e.g., "stra-stri-stru", "ba-gu-di"). Bohland et al. [16] found an effect of sequence complexity in the caudate in addition to other speech areas, but did not find an effect of syllable complexity in the caudate. This provides evidence for the role of the caudate in sequencing speech. It also motivates our hypothesis that the caudate will show a phoneme rate effect because the caudate may need to change its activation to scale with the rate of speech sequencing. Our study is different than Bohland and Guenther [15] because we will explicitly be analyzing the rate of sequencing, and we will do so by directly quantifying the rate of speech from participants in the scanner.

Another element of our study different from all published, prior work is the use of a state-of-the-art imaging technique, simultaneous multi-slice (SMS) acquisition [116]. SMS can enhance spatial and temporal resolution without sacrificing signal to noise ratio up to certain limits; a breakthrough that is as close to a "free lunch" as possible. This may enable greater sensitivity for detecting and localizing changes in brain activation.

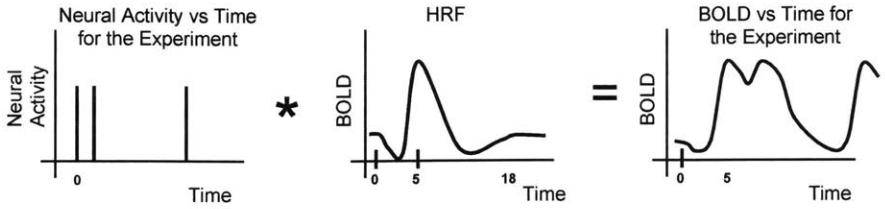


Figure 3.3: The BOLD response measured during an experiment can be well modeled as a linear convolution of the neural activity during the experiment convolved with the hemodynamic response function (HRF). This relationship forms the basis of the General Linear Model framework. Figure after [70].

■ 3.1.5 Neuroimaging of Speech Production

The Functional MRI Scan

Functional MRI (fMRI) measures the blood oxygen level dependent (BOLD) signal. The BOLD signal is the change in the amount of oxygenated blood in an imaging voxel. When neural activity occurs, a delayed hemodynamic response (a change in blood flow) follows in which an increase of oxygenated blood perfuses the region of neural activity. The hemodynamic response function (HRF) is the BOLD signal corresponding to a single, co-ordinated neural event ("single" does not mean single neuron but the simultaneous activity of many co-located neurons). This perfusion signal follows a curve approximately modeled by a double gamma function with a peak amplitude approximately five seconds after the neural activity which caused it to occur (see center panel of Figure 3.3). The exact BOLD HRF varies in shape depending on the brain region and among people. [70].

The measured BOLD signal during an experiment is approximately a linear convolution of the neural activity with the HRF. The general linear model analysis framework is founded upon this linear system approximation schematically shown in Figure 3.3. Given a measured BOLD response and knowledge of the experiment, the GLM determines at each voxel whether the BOLD response can be attributed to the experimental intervention (e.g., Is the measured BOLD signal due to the subject speaking an emotional sentence, due to unknown effects such as random noise, or due to a motion artifact?).

Because fMRI measures changes in blood oxygenation instead of electrical activity, the electrical neural activity in the cortex and the electrical activity due to muscle activation for overt speech production do not contaminate the fMRI signal. However, the BOLD HRF has a low signal to noise ratio, so fMRI relies on signal integration across brain slices and across brain volumes to improve signal detectability at the cost

of spatial and temporal resolution.

A brain volume is a single snapshot of the entire brain at one point in time. However, the entire brain is not imaged simultaneously, only approximately simultaneously (e.g over the course of one second). The time during which the brain is imaged is the acquisition time (T_A). Acquisition time is like the exposure time of a camera. When a subject speaks, the small tremors due to speaking physically change the segment of brain located inside a particular imaging voxel. Each voxel size is on the order of 2 mm by 2 mm by 2 mm. Consequently, much as a picture is blurred if the object moves, a fMRI brain volume will be blurred because the brain is moved while speaking during acquisition. Physical movement of brain regions between voxels not only distorts the particular brain volume during which movement occurred, but that same brain volume now no longer aligns voxel to voxel with previous and subsequent brain volumes.

Because movement during MRI acquisition introduces signal artifacts, it may seem that overt speech production cannot be studied by MRI. However, the delay between the neural activity that occurs during speech production and the HRF peak which is actually measured means that the brain can be scanned between speech productions. Acquiring brain volumes periodically with silent pauses between scans is called sparse imaging. Sparse imaging avoids artifact motion within a scan volume while still capturing the HRF response. Sparse imaging has the additional benefit of allowing us to cleanly record subject speech with a microphone. The speech will be uncontaminated by MRI scanner noise because the brain volume is not being acquired while the subject is speaking. Figure 3.4 in the next section shows an example of the sparse imaging timeline.

An important limitation with fMRI is its poor temporal resolution of neural activity. fMRI temporal resolution is on the order of seconds, except in specially designed chronometric fMRI protocols. The poor temporal resolution is due partly to the convolution of a 1 ms neural spiking event with an approximately 15 second HRF function. It is also due to the sparse scanning paradigm. MRI volumes take approximately one full second for collection of all slices for a brain, and each brain volume may be collected every three or more seconds during a sparse acquisition paradigm. Therefore, brain activity, which is a continuous phenomenon, is only being sampled discretely every four seconds.

The General Linear Model

The primary form of analysis of fMRI data uses the general linear model [70]. The general linear model (GLM) is at its core a multiple linear regression equation that attempts to explain measured brain activations at each location (voxel) in the brain in terms of the experimental intervention and other factors such as random noise. With the GLM, brain locations can be identified whose activity depends on such factors as speaking or not speaking, phoneme rate, sentence type, and depression severity. For a detailed overview of the GLM, see Appendix A.

fMRI analysis takes place at the subject level and across subjects. Analysis at the subject level is called level one or L_1 analysis. Analysis across subjects is called level two

or group level analysis. The results from each of the subjects are aggregated to determine an average subject's brain regions that are related to the regressors. Group level analysis also allows comparing subjects based on an attribute like depression severity to determine if brain activations for one group (depressed) are different than for another group (control).

3.2 The fMRI Task and General Linear Model Analysis

This section describes our experimental setup and analysis using the general linear model (GLM). Our goal was to determine if and where regions of the brain are sensitive to phoneme rate and magnitude of valence, and our hypothesis is that these regions will be in the caudate of the basal ganglia when the experimental factor is phoneme rate.

3.2.1 Methods

Participants

Subjects were recruited through email, website (voicesurvey.mit.edu), and fliers that advertised an opportunity to participate in a voice and depression study. All subjects were at least 18 years old. Subjects with hypothyroidism were excluded to avoid the potential confound of a metabolic factor of depression (but see [36] for why this may not have been necessary). Also excluded were subjects with psychosis as assessed by the Yale Prime Schizophrenia survey [88, 89] and dementia as assessed by the Mini Cog Dementia survey (www.alz.org) [17, 18]. Subjects that self-reported as depressed but failed to score a Beck Depression Inventory II level of at least 14 were also excluded as were subjects who may have had bipolar disorder as evidenced by a score greater than or equal to 22 on the bipolar disorder self test [72].

Medications were noted for controls and depressed subjects but were not used in this analysis. To assess whether medication might induce speech slurring, and therefore confound our results, we queried ehealthme.com which aggregates reports of medication side effects from the FDA. Upon review with this resource of several depression medications such as Zoloft, Lexopro, Atarax, Cymbalta, Wellbutrin, and Celexea, we found "slurred speech" was reported less than two percent of the time. Therefore, a drug speech interaction cannot be ruled out but is unlikely.

Study data were collected and managed using REDCap electronic data capture tools hosted at MIT [64, 97]. All subjects gave written consent, were financially compensated for participation, and the study had MIT Institutional Review Board approval.

Protocol

MRI compatible microphones were used to record overtly produced, implicitly emotional sentences from a subset of a corpus of sentences from Russ et al. [113]. Each sentence in the corpus was rated by participants in the Russ et al. [113] study on its appropriateness for happy, sad, angry, or fearful contexts on a 1 to 10 scale. Russ and

colleagues categorized sentences, and we selected sentences from the happy, sad, and neutral categories for presentation.

Furthermore, we assigned a valence to each sentence by transforming the mean ratings for it. We computed valence as

$$\text{valence} = \frac{\text{avg. happy score} - \text{avg. sad score}}{9}. \tag{3.1}$$

This normalized the valence to +/- 1 with +1 corresponding to "sounds very appropriate for happy" and -1 corresponding to "sounds very appropriate for sad". Zero equates to neutral valence. We compute strength of the valence as the absolute value of the valence.

The majority of subjects had two runs of 48 brain volumes each (Run 1 = 12 sad, 12 neutral, 13 happy, 11 null; Run 2 = 12 sad, 12 neutral, 12 happy, 12 null). A null trial is a trial with no speaking. Null trials are needed to identify brain regions that contribute to speaking vs not speaking. Intuitively, if all brain measurements were of the subject speaking and there were no brain measurements of not speaking to compare against, then we could not identify regions that were specific to speaking. The slight imbalance between happy and null trials in Run 1 was due to a programming error not discovered until analysis. Subjects with anomalous runs because of presentation software problems were noted and processed accordingly. Stimuli were presented via custom scripts written in Psychopy [101].

For reasons discussed in 3.1.5, we used a sparse imaging paradigm. A schematic of the sparse imaging process is shown in Figure 3.4. The repetition interval is the total duration of time that includes both the brain acquisition ("MRI") and the sentence production ("SPEAK"). This time is 4 seconds and is abbreviated as T_R. The acquisition time is the total time the brain is being imaged within one T_R. The acquisition time was 1.1 seconds and is abbreviated as T_A.

Subjects were presented with the sentence they would say and then cued to read the sentence by a change in letter coloring from white to green. The sentence was presented during the 1.1 second acquisition time (T_A) marked by the red "MRI" bar in the figure. The remaining 2.9 seconds marked by the blue "SPEAK" bar were used for repetition of the sentence.

fMRI Acquisition Parameters and Analysis Software

All scanning was performed at the Martinos Imaging Center at MIT with a 3 Tesla Siemens Trio scanner and 32 channel head coil. A T_1 weighted (4 multi echo MPRAGE [135]) scan was collected for registration of the functional data and FreeSurfer surface reconstruction. Time permitting, a T_2 was also collected. Functional images were acquired with a T_2^* weighted blood oxygenation level dependent (BOLD) echo planar SMS 5 sequence [116]. T_1 and T_2^* acquisition parameters are in Table 3.1. Resting, diffusion, and seven other functional tasks, time permitting, were also collected but not part of this analysis. Appendix B lists the full protocol.

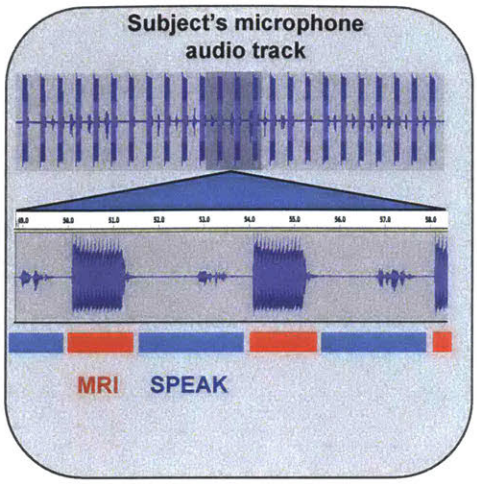


Figure 3.4: Sparse imaging protocol visualized by observing the recorded audio track. Periods of speaking (low amplitude) alternate with periods of MRI noise during the brain scan (high amplitude). The acquisition time is equal to the duration of the red bar, and the repetition time is equal to the duration of the red and blue bars.

The data were analyzed using Nipype workflows [55] using FSL [122], ANTS [7], and FreeSurfer [44]. Analysis was volume based, as opposed to surface based, so region of interest annotations on surface based renderings are from using volumetric atlas queries (FSLview with the Harvard-Oxford cortical and subcortical atlases [38, 65]). Surface renderings are a visualization aid only, and while we report many regions of activity, our verbal description is not exhaustive. Readers should realize that the actual brain volumes themselves are needed to answer specific questions about regional activity.

General Linear Model Analysis

We use the general linear model framework to analyze the fMRI data at the subject level (L_1 analysis) and at the group level [70]. For the L_1 analysis for each subject, the design matrix included a binary speaking/not speaking regressor, a speech rate regressor, and valence regressor. Following Mumford et al. [92], the speaking rate and valence variables are demeaned in order to orthogonalize them relative to the speech/no speech variable. Outlier volumes, as detected by the artifact detection software, and six motion regressors (3 translation, 3 rotation) complete the design matrix.

We report results using the articulation rate as our measure of speech rate as opposed to speaking rate. Recall, speaking rate is defined as number of phonemes per sentence, including silence, divided by the total duration of the sentence, including the silence within the sentence. By contrast, articulation rate is the number of phonemes per sentence, excluding silence, divided by the non-silent duration of the sentence. While

Table 3.1: Structural and functional imaging acquisition parameters. T_R= Repetition time, T_A= acquisition time, T_E= echo time, T_I= inversion recovery time. See Huettel et al. [70] for definitions.

Scan	Total Time (mm:ss)	T_R (ms)	T_A (ms)	T_E (ms)	T_I	Flip Angle (degrees)	Voxel Size (mm)	Field of View (mm)	Matrix Size (voxels)
T_1	7:07	2530	n.a.	(1.64, 3.5, 5.36, 7.22)	1400	7	(1, 1, 1)	(256, 256)	(256, 256, 176, 4)
BOLD (T_2^*)	3:33	4000	1090	30	n.a.	90	(2, 2, 2)	(256, 256)	(108, 108, 65)
T_2	6:38	6000	n.a.	454	2100	120	(1, 1, 1)	(256, 256)	(256, 256, 176)

a concept of hierarchical speech rate control might intuitively feature speech rate as opposed to articulation rate, we use articulation rate as a practical substitute because of the difficulties in automatic phoneme detection in the MRI scanner environment. In about ninety percent of the sentences, the two measures are identical, which is reasonable given that within a single 1.5 second sentence, a person likely is not going to have substantial pauses. However, in ten percent of the cases, the two measures differ, and upon inspection the difference can quite dramatic. When the underlying transcript is examined, it became apparent that the phoneme recognizer was declaring large sections of the sentence as silence, confusing speech with the background noise (e.g., helium pump and ventilation systems) of the scanner environment. This was artificially and incorrectly decreasing speech rate, but would have no effect on articulation rate since these silences would be excluded.

The output from the L_1 analysis is a set of brain activation maps per subject. Each brain activation map is called a contrast because it shows the difference in brain activity between two conditions. A simple condition is speaking and a second simple condition is not speaking. The speak vs no speak contrast shows brain regions with activity significantly different from not speaking (either more relative activation or less). We define baseline condition as not speaking. We report three contrasts corresponding to the significance of the three task based regressors: speaking different than baseline, articulation rate different than baseline, and magnitude of valence different than baseline.

We combined data from all subjects at the group level for each of the L_1 contrasts. The combination occurs in two ways (also called contrasts). Contrast 1 is the task effect. All the subjects are aggregated to test for common brain activation patterns. Contrast 2 is the effect of depression. Is there a difference in brain activation between the depressed and control groups? Group level results are shown in an MNI 152 coordinate system using Mayavi (http://docs.enthought.com/mayavi/mayavi/) and the Conte brain template [137].

We correct for multiple comparisons using family wise error rate (the probability of one or more false positive voxels [62]) and cluster level thresholding. First, we choose a voxel level threshold for significance (a.k.a. cluster detection threshold abbreviated CDT), and binarize the statistical brain map Then, based, on the connectivity and spatial extent of the detected voxels and a specified family wise error rate, a cluster level threshold is determined. Only voxels within clusters that have a cluster p-value less than a threshold are declared statistically significant. We use a voxel level threshold of $p = 0.01$ ($z = 2.3$). We then use a family wise error rate on the clusters of p=0.05. We use the FSL "cluster" method to perform these computations.

■ 3.2.2 Results

We present results from the general linear model analysis beginning with a description of the analyzed dataset, verifying the presence of the speech network, and then showing the contrasts of relevance to our hypothesis: that there is a depression by articulation

rate interaction in the caudate[1].

Dataset Characteristics

Due to the complex nature of the experiment, technical problems did impact subjects available for final analysis. At each stage of the process, some subjects were excluded. We graphically show this loss of useable data in Figure 3.6[2].

To understand the flow and subsequent loss of subjects through the analysis pipeline, we tap into the pipeline at several points. These points are: received an MRI (control: 25, depressed: 26, total: 51) T_1 was successfully collected and reconstructed (control: 24, depressed: 25, total: 49), fMRI data exists for the task of interest (control: 25, depressed: 26, total: 51), audio successfully recorded and processed for the task of interest (control: 23, depressed: 25, total: 48), input to the L_1 processing chain (control: 22, depressed: 24, total: 46), output from the L_1 processing chain after quality assurance (input to group level analysis) (control: 16, depressed: 20, total: 36).

The input to the L_1 processing chain is the four way intersection of successfully recording a fMRI, a T_1, audio, and the task.

The quality assurance step involved a behavioral check to make sure null trials were present for the paradigm[3], and a visual inspection of the L_1 contrasts in order to check for a banding artifact. The banding artifact is believed to occur because of subject motion and the SMS sequence. The SMS sequence creates different steady state excitations across a stationary brain that should be the same for all brain scans. If the subject moves between brain volumes, the steady state excitation that was expected from a particular location has moved to a new location. The mismatch between expected and actual excitation may yield banding. A representative example of banding is shown in Figure 3.5. Two subjects had no null trials for either run (986 and 983), and one subject had no null trial for run 2 and banding for run 1 (862). Other subjects were excluded because all collected runs had banding artifacts.

The demographics of subjects who entered the final group level analysis are shown in Tables 3.3 and 3.2. All controls were right handed. Two depressed females were left handed and one depressed male was left handed. We report these demographics because fMRI activations in general are known to be different with age, sex, and handedness. Therefore, it is important to have as matched a set of controls and depressed subjects as possible in these attributes otherwise the differences that might be detected could be attributed to a characteristic other than depression. We report the median and interquartile range as summary statistics that are robust to outlier subjects. The median is the 50^{th} percentile of a distribution, and the interquartile range (IQR) is the difference between the 75^{th} and 25^{th} percentile of a distribution.

[1] All results are for level one model 455, group level model l2_binbdi_m455_task_onlyNull_orthbdi, and output folder l2_binbdi_m455_task_onlyNull_orthbdi_l1output_fs6_aggTrue_455

[2] cobidas_demographics.ipynb

[3] make_scanner_null_matrix.ipynb

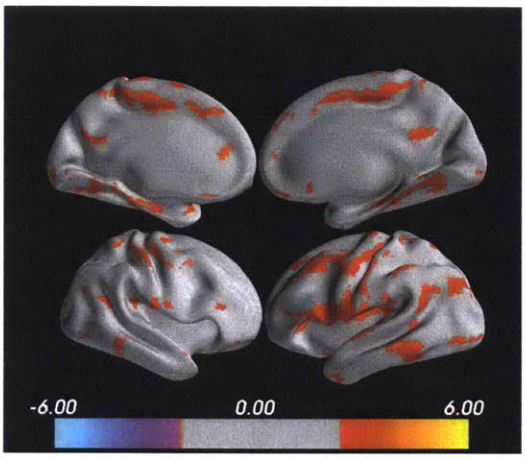


Figure 3.5: A representative example of the banding artifact that was grounds for subject exclusion. Banding is believed to occur from subject motion during a run that uses the SMS sequence.

Table 3.2: Number of subjects (Count) analyzed at the group level in the dataset, and the Beck Depression Inventory statistics for the subjects. IQR: Interquartile range.

		Count	Mean	Std. Dev.	Median	IQR	Min.	Max.
	Sex							
Control	F	8	2.8	2.8	1.5	4.50	0	7
	M	8	2.6	2.5	2.5	3.25	0	7
Depressed	F	10	25.1	10.0	23.5	14.75	14	44
	M	10	30.8	5.7	30.0	9.75	23	38

Table 3.3: Number of subjects (Count) analyzed at the group level in the dataset, and the age (years) statistics for the subjects. IQR: Interquartile range.

		Count	Mean	Std. Dev.	Median	IQR	Min.	Max.
	Sex							
Control	F	8	28.5	8.0	27.5	5.0	19	46
	M	8	28.4	9.9	25.0	7.0	20	51
Depressed	F	10	28.9	9.5	27.5	8.5	18	52
	M	10	31.9	16.2	23.5	18.5	18	65

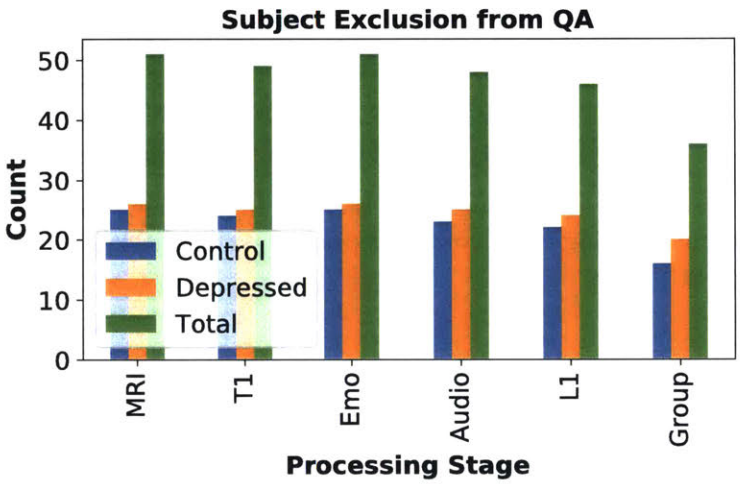


Figure 3.6: Subject exclusion based on quality assurance and data availability. We show by subgroup (depressed and control) and by total subject count the number of subjects with data at various processing stages. Starting from all subjects who received an MRI ("MRI") and completed the emotional sentences task ("Emo"), we show how many subjects had successful data collections and processing. "T_1" tallies the subjects that had a successful T_1 reconstruction, "Audio" tallies the number of subjects with successful audio recording, "L_1" tallies the subjects with audio, a T_1, and a fMRI, and "group" shows the number of subjects from the Level-1 analysis who did not have banding artifacts and could be analyzed at the group level.

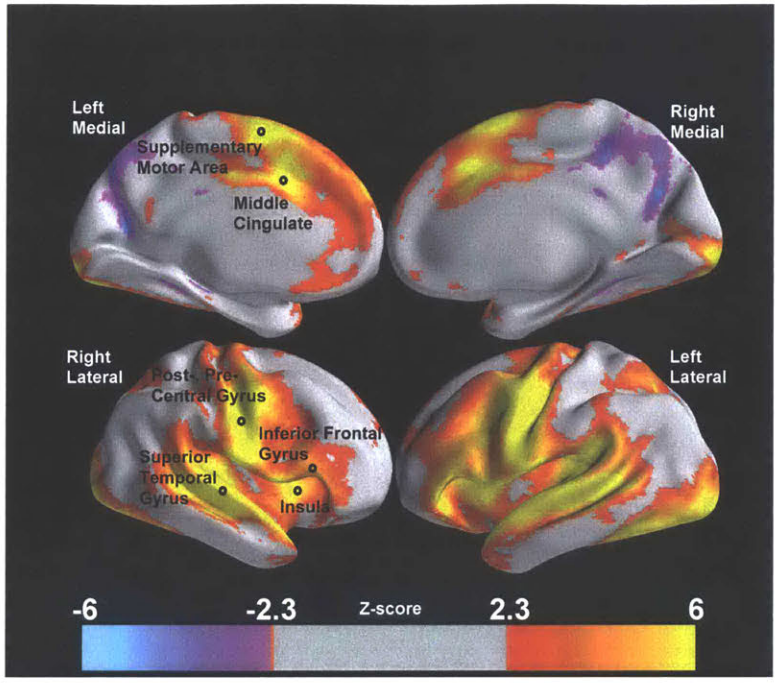


Figure 3.7: A robust group average speech network is recovered, medial (top) and lateral (bottom) views. L_1 contrast: task > baseline, Group: all subjects.

Main Effect: Speaking vs Not Speaking

We demonstrate robust activation of the speech network across individuals and at the group level in Figure 3.7. Bilateral activation is present on the medial surface in the top figure covering SMA and preSMA. IFG, STG, and MC are present on both lateral surfaces, and left hemisphere (pictured on the right hand side of Figure 3.7) shows a more distributed set of activations consistent speech having a left lateralized bias. The coronal section in Figure 3.8 shows putamen and auditory cortex activity. Activations include supplementary motor area, basal ganglia, motor cortex, and superior temporal cortex (auditory cortex), which is consistent with activations in overt speech production tasks [60][4]

Main Effect: Valence Intensity

Figure 3.9 shows the main effect of the strength of valence. There is significant activation in the left medial view including superior frontal gyrus and frontal pole, anterior cingulate cortex, and subcallosal cortex. These regions are present to a lesser extent in the right hemisphere. In both hemispheres laterally there are strong activations running

[4]Group plots from plotFlat3d/wrapper_gp.ipynb

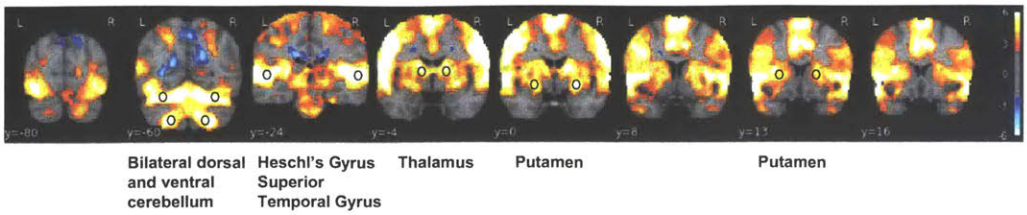


Figure 3.8: A robust group average speech network is recovered, coronal cross sections. L_1 contrast: task > baseline, Group: all subjects.

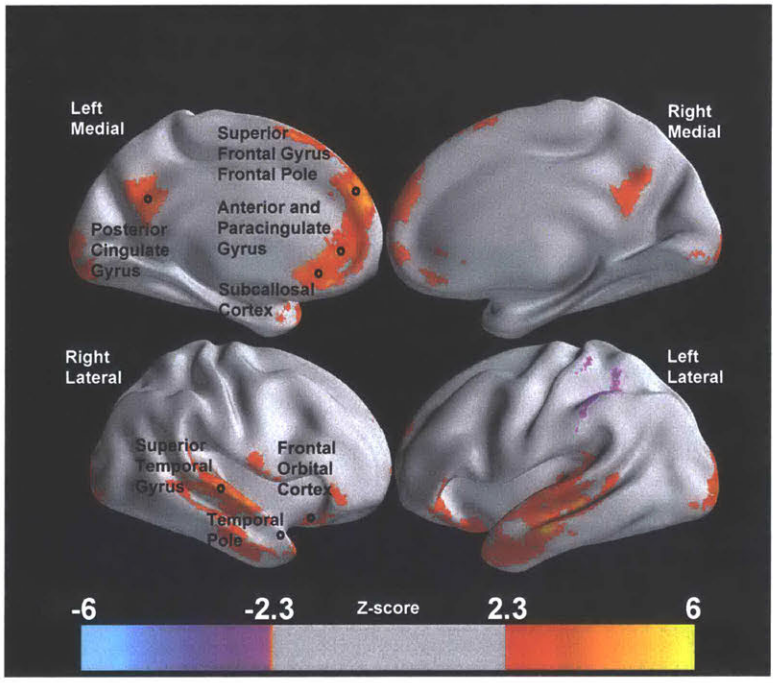


Figure 3.9: The main effect of the absolute value of the valence reveals extensive limbic cortex activation across all subjects. L_1 contrast: absolute value of valence > baseline, Group: all subjects.

from the temporal pole along the superior temporal gyrus and middle temporal gyrus. Figure 3.10 shows these regions as well as the amygdala and frontal orbital cortex and insular cortex. A set of significantly negative activations are present bilaterally in the superior parietal lobule, supramarginal gyrus, and angular gyrus (not shown).

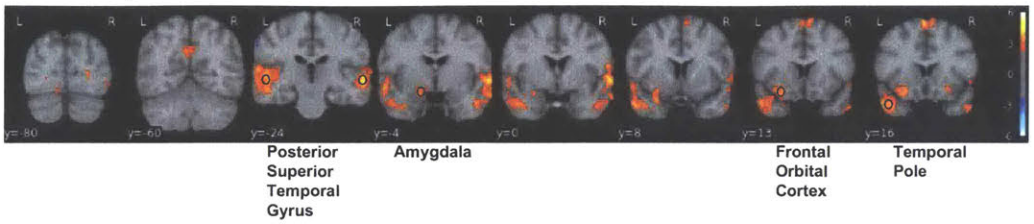


Figure 3.10: Coronal view of limbic activations with sentence valence. L_1 contrast: absolute value of valence > baseline, Group: all subjects.

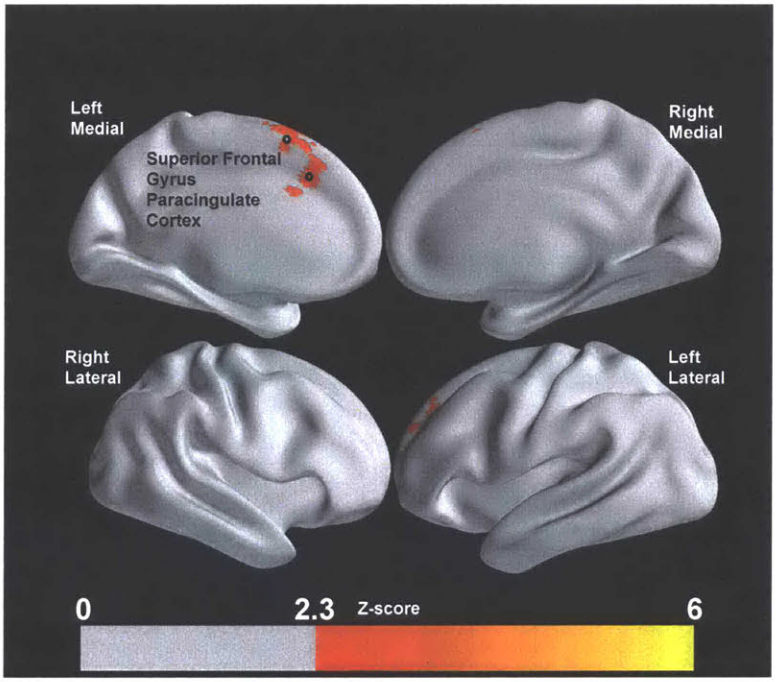


Figure 3.11: The group effect of controls vs depressed for the absolute value of valence reveals increased activation in the superior frontal gyrus and paracingulate cortex for controls relative to depressed. L_1 contrast: absolute value of valence > baseline, Group: controls > depressed.

Controls vs Depressed Effect: Valence Intensity

Figure 3.11 shows the cortical activations for the controls greater than depressed contrast for the absolute value of the valence. There is more activation in the left superior frontal gyrus and left paracingulate cortex for controls than depressed as well as the left frontal pole (not visible in the figures).

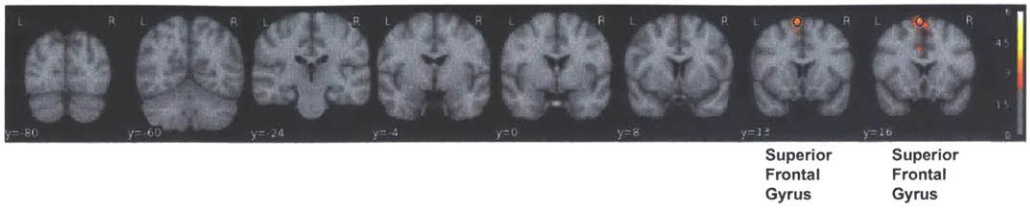


Figure 3.12: Coronal view of limbic activations with sentence valence. L_1 contrast: absolute value of valence > baseline, Group: controls > depressed.

Main Effect: Articulation Rate

We see some decreased activation in the frontal pole and superior frontal gyrus. There is some decreased activation in the post central gyrus, and to a lesser extent, the precentral gyrus. However, we caution interpreting any of these results because they are driven by a single subject, 961. Without subject 961, essentially only a small decreased activation in the frontal pole is still statistically significant.

Controls vs Depressed Effect: Articulation Rate

We do not see any significant activation for the controls vs depressed contrast with respect to articulation rate.

Controls vs Depressed Effect: Speaking vs Not Speaking

We see in Figure 3.13 a depression by speaking (not articulation rate) interaction at the group level[5]. There is more activity in controls than depressed when speaking vs not speaking. This activity is left lateralized and extends through the insular cortex and the putamen. There is also bilateral activation in the precuneus and lateral superior occipital cortex. This activation cluster can be seen clearly with axial slices in Figure 3.14.

In Figure 3.15 we create a subject scatter plot of the effect size in the left putamen and other regions using anatomically defined regions of interest (ROI) from each subject's Freesurfer parcellation. The mean effect size is the contrast of parameter estimate (COPE) for the task regressor. Intuitively, the COPE is proportional to how strongly the brain activation timeseries for the subject is correlated with the regressor. We visualize the COPE vs BDI to further look for a trend vs a categorical difference between the depressed and control populations. We present the Spearman correlations (uncorrected) in accompanying Table 3.4.

[5] cope_vs_bdi_roi_fssummary.ipynb

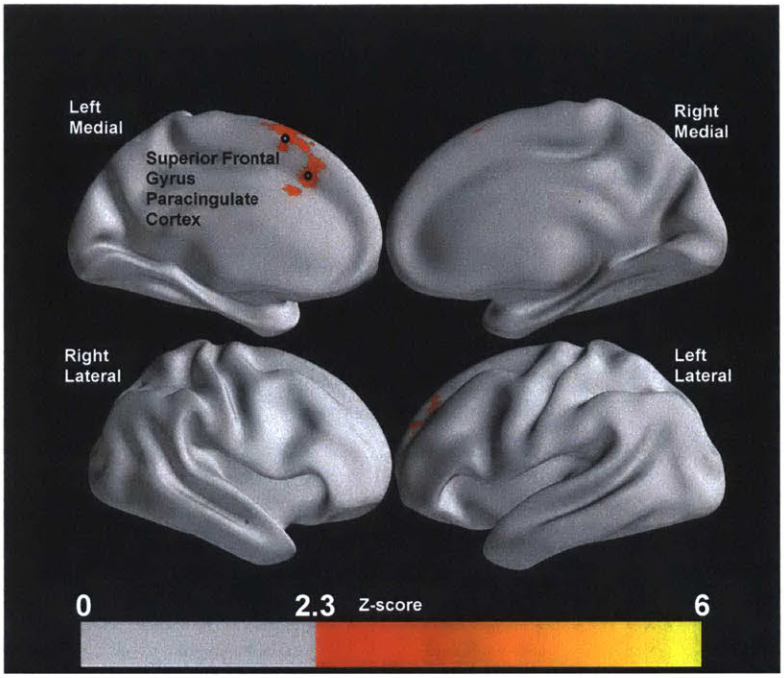


Figure 3.13: The group effect of controls vs depressed for the absolute value of valence reveals increased activation in the superior frontal gyrus and paracingulate cortex for controls relative to depressed. L_1 contrast: absolute value of valence > baseline, Group: controls > depressed.

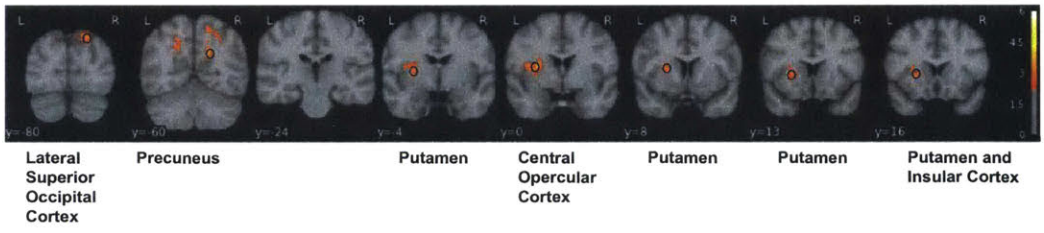


Figure 3.14: There is a strong putamen and anterior insula activation difference between controls and depressed subjects when speaking. L_1 contrast: task > baseline, Group: controls > depressed.

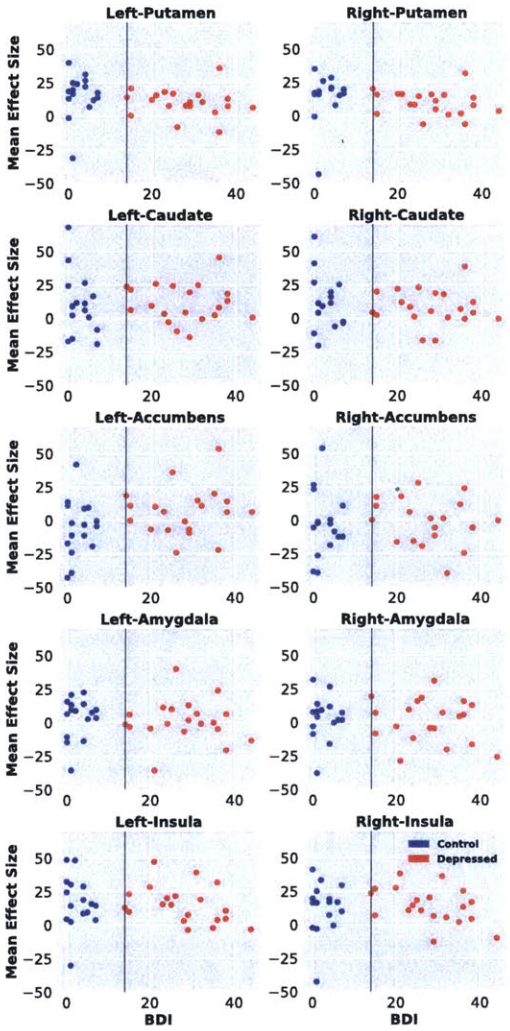


Figure 3.15: The subject level effect sizes within the left putamen and other regions that were used in the group level contrast of Figure 3.13. The vertical line corresponds to the BDI threshold of 14 for classification as depressed, and the effect sizes differ between the groups. This plot visually shows the interaction of depression with speaking because there is a linear relationship with non-zero slope between depression severity and effect size. L_1 contrast: task > baseline, Group: controls > depressed.

Table 3.4: Spearman correlations of COPE values with BDI for various regions of interest.

ROI	Spearman Corr.	p value
Left-Putamen	-0.38	0.02
Right-Putamen	-0.43	0.01
Left-Caudate	-0.05	0.77
Right-Caudate	-0.17	0.32
Left-Accumbens	0.29	0.08
Right-Accumbens	0.02	0.92
Left-Amygdala	-0.13	0.43
Right-Amygdala	-0.11	0.54
Left-Insula	-0.25	0.14
Right-Insula	-0.13	0.44

■ 3.2.3 Discussion

Recall the research hypothesis: depression interacts with speaking rate in the caudate and depression interacts with valence in the caudate via an amygdala connection from the ventral amygdalofugal pathway. While we did see main effect of emotional valence in subcortical and limbic regions, we did not see an interaction effect with depression. The main effect is consistent with the Pichon and Kell [106] study.

While we did not see a speech rate interaction with depression, we did see a speaking interaction with depression that was in the vicinity of our hypothesized region. Instead of the caudate being robustly implicated we saw the insula and putamen. The insula and putamen are both involved in speech production, and the insula also is part of limbic cortex.

Therefore we were unable to find evidence to support our hypothesis as stated. Instead, we conclude from this study that we may not have had sufficient power to identify speaking rate as a behavioral variable but were able to identify speech itself as being modulated by depression. Depression severity causes a dampening of brain responses with respect to speaking in the putamen. This could be reflective of vowel space reduction observed in other speech studies of depression [114] because the putamen is implicated in the execution loop of speech production [16]. We saw that the differential activation of the basal ganglia, perhaps by the insula, may play a role in creating speech biomarkers of depression.

One might ask whether a strong interactive effect of speaking rate and depression was not detected because of the experimental setup or more fundamental limitations of fMRI. To investigate this, we performed a two-sided Welch's t-test on the articulation rate between depressed and control subjects for their utterances (first aggregating

Table 3.5: Articulation rate (phns/sec) for read passages and emotional sentences for depressed and control subjects. Depressed > controls.

Protocol	Group	Mean	Std. Dev.	t-stat	p-value
Rainbow	Control	11.3	0.63	-2.73	0.01
	Depressed	10.6	0.76		
Grandfather	Control	10.8	0.82	-1.72	0.10
	Depressed	10.2	0.81		
Caterpillar	Control	11.2	0.62	-1.56	0.13
	Depressed	10.7	0.94		
Emotional Sent.	Control	11.1	0.60	0.56	0.57
(inside MRI)	Depressed	11.2	0.74		

articulation rates per subject) and compared the results to the read passage subjects[6]. While there is only a statistically significant difference between groups on the rainbow passage, in all cases of the read passages, there is mean difference between groups with the controls having a greater articulation rate than depressed. Inside the MRI, this trend actually reverse by a small amount and the p value for rejecting the null hypothesis that the two groups have the same mean becomes substantially larger than the p values in the read passages. In other words, there was no real behavioral effect to be observed inside the MRI.

However, despite there not being a behavioral effect, that does not rule out the possibility of a neural difference that leads to similar behavioral effects. This may be why we saw an overall difference in putamen activity with depression rather than a specific interaction with phoneme rate.

As we are still interested in understanding phoneme rate in the brain with a view towards using this knowledge to expand models from which features could be extracted, we revisit the neuroimaging modality and experimental paradigm. From smallest to greatest change in approach, we can consider altering the speech protocol, the scan protocol, and the neuroimaging modality. The simplest approach would be to change out the emotional sentences for sentences that feature the most differentiating phonemes that were discovered for the read passages. However, as we saw with the passages, the phonemes are variable across passages, so the sentences should be taken from the passage as the stimuli sentences.

Next, we might change the scan protocol from sparse imaging to continuous imaging while the subjects read the full rainbow passage without artificially segmenting the rainbow passage into sentences. While superficially this would recreate the read passage experiment inside the scanner, the continuous MRI noise would change the speech production experience, and this noise would also be a confound with the participant's speech in the audio recording (but see Bresch et al. [19] for one way to remove MRI

[6] mit_emo_behavAcoust.ipynb

noise from speech). On the other hand, the rainbow passage is not particularly emotive, so the ideal would be to read an emotional story of some kind rather than a neutral passage.

With the dataset as acquired, we can still analyze the resting state and diffusion data for indications of neural correlates that could give rise to behavioral differences. In the resting state data, we might establish a network of connections and test the predictive power of the network against the articulation rate of individual phonemes or a fused phoneme feature (where the fused phoneme feature might be derived as in Chapter 2). A similar prediction procedure could be followed with white matter tracts especially the dominant ones associated with the speech network. Of particular interest would be the temporo-frontal extreme capsule fasiculus as it connects the temporal lobe to inferior frontal gyrus and therefore connects part of the limbic system to the speech network [104].

The most extreme change would be a neurological intervention within the basal ganglia (for a review see Skodda [119]). However, to foreshadow the modeling chapter, we observe that speech rate is controlled not just at the phoneme level but also at multiple levels of speech production. Furthermore, speech rate is not a "one shot" phenomenon in which a rate is entirely planned and then execution proceeds. Speech rate is a dynamic process that continues during production. As summarized in Skodda and Schlegel [120], the reported alterations are not confined to speech rate alone, so it would be difficult to draw a conclusion about the sufficiency of a stimulated region for affecting speech rate though necessity could be established.

■ 3.3 Connectivity Analysis

The general linear model analysis we conducted in the previous section failed to localize a main effect of speech rate, and it failed to detect any subcortical interaction between speech rate and depression. While such effects may not be detectable with the experimental paradigm due to small effect sizes, we still believe there should be some representation of rate in the brain and some form of interaction of rate with depression. Instead of these effects being localized to particular points, these effects may be encoded by the interaction among several points in the brain each of which is too weak to be detected by itself through the general linear model. To approach analysis of speech rate, valence, and depression from a network perspective, and to include prior knowledge, we turn to dynamic causal modeling.

Dynamic causal modeling (DCM) [51] is an analysis technique for determining connections between brain regions using mechanistic models of neural dynamics and hemodynamics. We have two goals for DCM both of which serve the theme of this thesis in developing biomarkers of depression from models of speech.

First, we can directly use the connection weights determined by DCM for each subject as features with which to classify depressed vs control subjects. If this is possible, then we will have shown that a network representation of brain activity during speaking

is encoding useful information. It is impractical to scale such a feature to worldwide use as everyone would need to have an fMRI performed. However, if the features did encode this predictive information, we would have a tool in the form of a subnetwork of the brain which was indicative of depression during speaking.

Second, we can ask the more general question of whether or not a DCM style model of the brain is a useful encapsulation of the underlying processes. We will use the connection weights determined by DCM in one fMRI run by a subject to predict the HRF responses in the second run of the fMRI task. If DCM is capturing intrinsic structure, the model trained on the first run should be able to predict HRF response during the second run, and vice versa.

■ 3.3.1 Background

We applied DCM to our data as a model based approach for understanding the connections between brain regions and how those connections might differ between depressed and control subjects with the intention of using these differences as predictive features of depression. Specifically, we focused on the connections within and between the speech and limbic networks and the modulation of the brain regions in these networks by the experimental task.

Dynamic Causal Modeling

DCM models the neural state of each node of the brain in the same way. DCM assigns to each node a neural state value, x that represents the amount of neural activity occurring in the node. This neural activity evolves with time according to a bilinear differential equation

$$\dot{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{x} + \sum_{j=1}^{m} \boldsymbol{u}_j \boldsymbol{B}^j \boldsymbol{x} + \boldsymbol{C}\boldsymbol{u}. \tag{3.2}$$

$\dot{\boldsymbol{x}}$ represents the time derivative of the neural state, and $\boldsymbol{u}$ represents experimental input (e.g. speech rate, sentence valence, or other experimentally measured or manipulated variable). $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are the connection matrices whose values are to be estimated. $\mathbf{A}$ is the static connections between nodes. However, these baseline connection strengths can potentially be modulated by experimental inputs. For example, the baseline connectivity between two brain regions might be large while resting, but when engaged in a task, the actual connectivity might be decreased. $\mathbf{B}^j$ describes how experimental input j might change the baseline connection strengths.

The change in neural state of any given node is the weighted sum of the node's current state and its neighbors where the weights are set by the $\mathbf{A}$ and $\mathbf{B}$ matrices plus the direct input from experimental variables ($\mathbf{Cu}$).

Neural state is not directly measured by fMRI, only the hemodynamic response is measured. Consequently, DCM includes an observation model in which the neural state is convolved with a hemodynamic response in order to create a predicted BOLD

response. The neural state equation with **A**, **B**, **C** estimated and the observation model comprise a complete model of the brain that translates from experimental activity into observed fMRI responses.

DCM takes as input the measured HRF timeseries for each region of interest, an *a priori* set of connections among the regions of interest (ROI), a set of stimuli time series, and an *a priori* set of connections from the stimuli to the regions of interest. Then, intuitively, DCM generates an HRF waveform for each ROI given the current model connections, the stimulus, and known neuronal and hemodynamic functional relationships. DCM compares the generated HRF to the measured HRF and updates the connectivity weights based on the error. DCM iteratively updates the connectivity weights in a cycle of generating the HRF and correcting the weights until convergence is reached.

The mathematical implementation of this process is slightly different than the given conceptual explanation [50]. DCM in the Statistical Parametric Mapping (SPM) MATLAB (The Mathworks, Natick, MA) software package [3] minimizes a quantity termed "free energy" which is related to how likely the data came from a set of model parameters. The optimization process accounts for model complexity, so a fully connected model (e.g. all brain regions are interconnected) will not automatically score better when compared to a more sparsely connected model [102].

Research Hypothesis

Our original research hypothesis was that phoneme rate interacts with depression in the caudate, and that this interaction is made possible by the amygdala's connection to the caudate. Therefore, we specify two models of neural connections (ie. two different **A** matrices) and will compare them to check for the amygdala to caudate connection. Note, we are not specifying the value of the connections, only their presence or absence. DCM will solve for the most likely values of the connections given the data.

■ 3.3.2 Methods

Model Specification

Figure 3.16 shows the connectivity matrix, **A**, in graphical form. Each yellow square is a connection from the brain region in the corresponding row to the brain region in the corresponding column. Each purple square is the absence of a connection. The presence or absence of a connection was determined from DIVA model connections and regions in Tourville and Guenther [128] as well as numerous other primary and secondary sources. In our literature search, if we found one connection from a node to a node, we also entered that connection for the reverse direction to create a symmetric matrix. The presence of cortical-subcortical loops and bidirectional cortical- cortical connections is an established feature of neural architecture that we use to justify creating a more flexible model than if connections were unidirectional.

We capture the difference between the two **A** matrices we will test by shading two

of the blocks light blue. These represent the connection between the left amygdala and the left caudate (consistent with our hypothesis), and the connection between the left accumbens and the left caudate. We include the left accumbens because the accumbens is a key part of the limbic system, and anatomically shares a border with the caudate. Therefore, we believe it is reasonable that this is another avenue by which the limbic system can influence the caudate. We refer to the **A** matrix in which these connections may exist as $\mathbf{A}_{\text{connected}}$, and the **A** matrix in which these connections are prohibited as $\mathbf{A}_{\text{disconnected}}$.

In addition to our two different **A** matrix initializations, we must initialize a **B** matrix of modulatory connections, and a **C** matrix of experimental inputs. As with the **A** matrix, these matrices only specify the possibility or not of a connection, not the connection weight. If a connection is not allowed, the disconnection is enforced.

Our **B** matrix is all zeros which represents the absence of the modulation of connection strengths between regions by the experiment. We believe that this is a reasonable first approximation that allows us to focus on the possible differences between the baseline **A** matrix networks between depressed and control subjects.

Our **C** matrix includes three inputs corresponding to the three experimental variables in our fMRI task: the presence or absence of speaking, the absolute value of the stimulus valence, and the articulation rate. The speech/no speech was connected as a driving input to SMA because of the SMA's role in speech initiation and to Broca's area (lh_vIFo) for its role in speech sound map retrieval. The valence modulator was connected to the anterior insula. The anterior insula was shown by our GLM analysis to be active for the main effect of task for emotional sentences. Furthermore, the anterior insula is also observed in the literature to be active for emotional processing under cognitive demand [105]. The phoneme based articulation rate was connected to the caudate.

Model Input Data

Hemodynamic response function (HRF) data was taken from the pre-processing output of a Level-1 workflow. As part of the pre-processing, the HRF data was realigned and highpass filtered, but not smoothed. The smoothing operation was omitted because smoothing would cause HRF activity to blur across region of interest boundaries. Instead, we parcellated the voxels in the brain, and took the median among the voxels within each parcel. The median operation was chosen over the mean as a summary statistics to avoid contamination by stray voxels that contained white matter. We used the motion compensation matrix (3 translation and 3 rotation regressors) and outliers found during the L1 pre-processing as confound factors in the DCM estimation. Data was analyzed with custom python Nipype scripts wrapping SPM 2012's dynamic causal modeling toolbox [3].

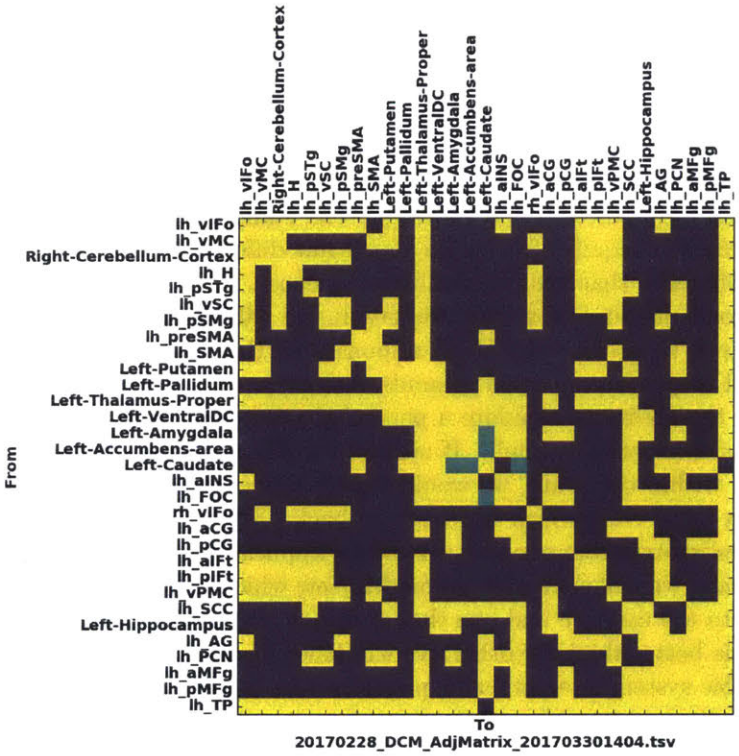


Figure 3.16: DCM **A** connectivity matrix. Intrinsic connections present in yellow, disconnections in dark blue. Hypothesized connections in light blue.

Evaluation

Traditional DCM model comparison uses a measure of model fit returned by the DCM fitting process and compares the ratio of fit between two models. If one model fit is substantially better than the other, the best fit model is accepted as significantly better than the other model.

However, our aim in modeling connections within the brain is to rigorously test how well the models are capturing a fundamental aspect of individual connectivity. Therefore, we will compare models in a cross validation- prediction framework. We will use the connectivity weights from each of the models as feature vectors and predict from those weights whether or not a subject is depressed. We will use leave one subject out cross validation and the same pre-processing and model construction as detailed in the previous chapter. If the connectivity matrix from one of the two models leads to better classification performance than the other when used as a feature vector, we will declare the corresponding model as the superior model.

Evaluation by prediction is a more rigorous standard than traditional evaluation methods because it requires not just a significant difference between the models but a significant difference that has practical consequences.

As a second measure of model evaluation, we will use parameters fit on one run for one subject to predict the HRF responses of the second run from the subject. This method of prediction again grounds evaluation of the model against a tangible application: the ability to simulate a person's brain response to new stimuli based on training data from other stimuli. If either or both models is capturing fundamental organization within the brain, we would expect this organization to be stable at least between two fMRI runs which are back to back. Therefore, the model connectivity weights, as estimated from one run, should be applicable to the second run.

In summary, we will compare two models, one which includes the presence of limbic connections to the caudate and one that does not. If the alternative model with these connections is better than the other, we will have garnered support for our hypothesis that the limbic system interacts with phoneme rate in the caudate.

As a practical matter of optimization, data is reduced using SPM's DCM dimensionality reduction to no more than eight functional nodes instead of the nearly three dozen nodes in the full connectivity graph.

■ 3.3.3 Results

Dataset Characteristics

As an introduction to what the DCM output looks like, we show Figure 3.17. We trained the model using HRF timeseries from this subject for all the regions we identified in the brain, and embedded within our model. After estimating the model coefficients for one run from the subject, we predicted the brain response for that same region for the other fMRI run performed by the subject and compare against the true brain response for that region. We visualize the predicted and actual waveforms by normalizing them

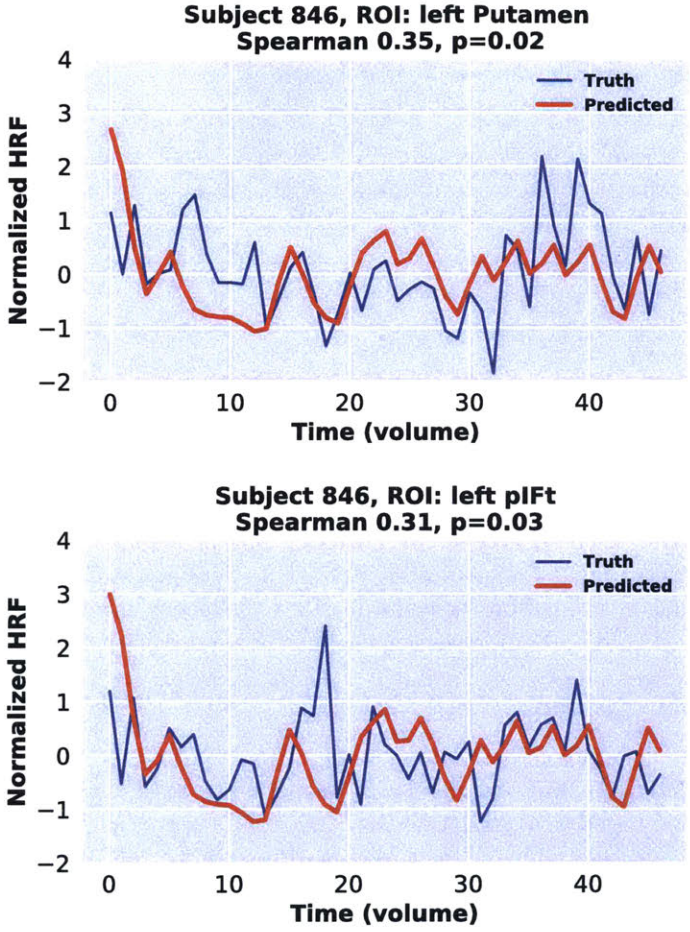


Figure 3.17: Two examples from subject 846 showing the predicted and true HRF waveforms in the left putamen (top) and inferior frontal gyrus, pars triangularis (bottom) for the first fMRI run. The predicted waveform is generated using connectivity weights estimated from the second fMRI run, but stimuli from the first fMRI run (i.e., true out-of-sample prediction). The estimated connectivity matrix is $\mathbf{A}_{\text{connected}}$. p values are uncorrected.

to have comparable means and variances, and we see good agreement. The Spearman correlation is greater than 0.3 for both regions[7].

Unfortunately, not all runs and brain regions showed such excellent predictive match. Furthermore, even when the the model is simply fit to the data (i.e., not out-of-run

[7] pack_dcm_pred.ipynb

Table 3.6: Receiver operating characteristic (ROC) area under the curve for classification performance using DCM **A** matrix weights for the connected and disconnected models. IQR: interquartile range.

		Mean	Std. Dev.	Median	IQR
Feature Set	Feature Subset				
$\mathbf{A}_{disconn.}$	weights	0.52	0.03	0.53	0.04
	all	0.49	0.03	0.48	0.03
$\mathbf{A}_{conn.}$	weights	0.51	0.03	0.50	0.03
	all	0.46	0.04	0.47	0.05

prediction), there is large variability in model fit as evaluated using the Spearman correlation between the fit and actual waveforms. For example, the left amygdala's waveform is able to be significantly predicted under thirty percent of the time whereas left Heschl's gyrus and left pre-SMA can be be significantly predicted over seventy percent of the time. These regions are part of a broader trend which shows difficulty in predicting limbic areas and less difficulty in predicting waveforms in sensorimotor regions. "Significant prediction" here means that the Spearman correlation coefficient between the predicted and actual waveforms meets a p <0.05 uncorrected threshold.

Model Selection by Classification Performance

We follow a shuffle-split classification protocol and machine learning pipeline as described in Section 2.2.4. We consider the **A** connection weights only as features (feature subset="weights"), and we consider all the connection weights as well their probabilities and the goodness-of-fit (GOF) as features (feature subset="all"). We evaluate performance using the receiving operating characteristic area under the curve (AUC). Recall, an area under the curve of 0.5 is chance, and 1.0 is perfect classification performance. Results are in Table 3.6[8].

We also remark that each run for a subject results in a set of DCM weights. If a subject had multiple runs, the weights are averaged. Unfortunately, not every subject in the general linear model analysis also had at least one run whose DCM weights converged. This could be because of a large number of motion artifact volumes that were removed for example. In total, this analysis used 33 subjects (15 control), and 56 fMRI runs.

The overall classification performance is in favor of the disconnected model provided only the weights are used as features in the classification pipeline, but the difference in ROC AUC performance between the two models is almost non-existent.

[8] pipe_sg_np.ipynb, pipe_fusion.ipynb, pack_dcm.ipynb

Model Selection by HRF Prediction

To summarize overall model goodness-of-fit (GOF) in order to compare between our "connected" DCM matrix and "disconnected" DCM matrix, we compute the average Spearman correlation across all ROI's and all runs for the cross-run prediction scenario. These include both the significantly correlated runs and the insignificantly correlated runs. We also compute this measure for the "fit" runs to provide a ceiling for how well we might expect to do in prediction. Finally, we break down the overall performance by control and depressed groups to explore whether one of these groups has more predictable or better fit brain responses than the other. Such a difference might be exploitable as a biomarker of depression. Again, because of the available data, these plots are over slightly different numbers of runs. The out-of-run predictions each used 73 runs and the within run fits each used 75 runs.

Figure 3.18 shows the four relevant cases[9]. As expected, overall performance is better under "fit" rather than "predicted" test cases. However, we do not observe any mean difference between the connected prediction model and the disconnected one as we saw when we used classification performance as our metric. Furthermore, we do not see a difference in model fit between depressed or control subjects. Additionally, the overall average goodness-of-fit, even in the case of within run fit, is still quite low with Spearman correlations barely greater than 0.3. In out-of-run-predictions, the GOF falls to below 0.1.

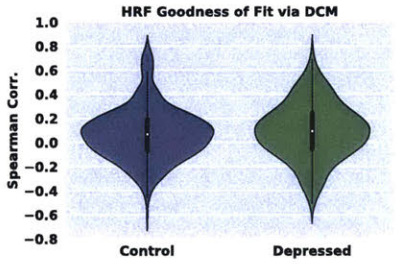
■ 3.3.4 Discussion

To summarize our findings from the DCM connectivity analysis, we were not able to differentiate between the two possible connection models based on the ability to perform cross run HRF prediction. Overall performance was highly variable across brain regions and between subjects. However, an interesting observation is that we were more often able to fit motor regions well than subcortical limbic regions. This may reflect how the experimental input from the speech task and the speaking rate drive a stereotypical response in sensorimotor regions. By contrast, the actual emotional state of the brain, as summarized by activity in amygdala, accumbens, temporal pole, and hippocampus can be quite variable and may depend on unmodeled connections and experimental stimuli outside of the valence of the sentence.

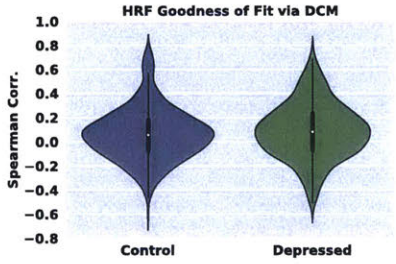
Some limitations of this work is the homogenization of the brain regions used within the parcellation. For example, the thalamus is known to be composed of several nuclei each of which themselves have distinct cortical and subcortical connectivity. However, these have been grouped together and simplified as a single entity for whom a single HRF timeseries is used to represent the activity of the node.

An open question is whether there is sufficient data to allow fitting of the DCM model. Assuming a dense graph, the DCM model has 31^2 connections whereas the timeseries for each region contributes 48 points (the number of brain volumes in a fMRI

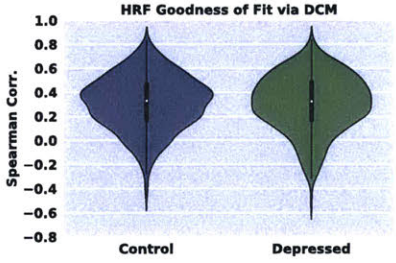
[9] pack_dcm_predict.ipynb



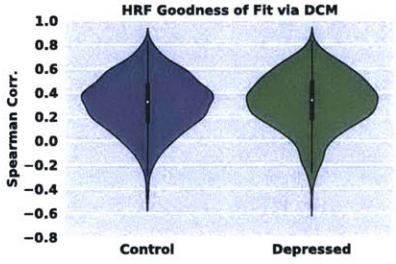
(a) $\mathbf{A}_{\text{conn.}}$: out-of-run prediction, GOF: 0.08 (0.23).



(b) $\mathbf{A}_{\text{disconn.}}$: out-of-run prediction, GOF: 0.08 (0.23).



(c) $\mathbf{A}_{\text{conn.}}$: within run fit, GOF: 0.31 (0.23).



(d) $\mathbf{A}_{\text{disconn.}}$: within run fit, GOF: 0.31 (0.23).

Figure 3.18: Comparison of the $\mathbf{A}_{\text{connected}}$ and $\mathbf{A}_{\text{disconnected}}$ matrices by considering the within run and out-of-run fits across all regions of interest for controls and depressed subjects. Goodness-of-fit (GOF) was evaluated by the Spearman correlation (mean, standard deviation) between the true and predicted runs.

run). As nodes are added, the model complexity grows with the square of the number of nodes, but the data support grows linearly. We mentioned in the methods that the optimization proceeds after dimensionality reduction is applied to eight nodes. However, we did attempt a brute force optimization without dimensionality reduction. After 24 hours of running, only a handful of subjects had reached convergence. Furthermore, when those subjects' HRFs were inspected, we observed a very strong low pass effect. It appeared that the algorithm was converging upon a simple sinusoid rather than predicting brain responses at each time point.

A possibility for investigating the data limitation of DCM exists in making a grossly reduced graph. For example, a two node graph, connecting the amygdala to the caudate or the anterior insula to the putamen (as we might do based on the results of the GLM analysis). With a simple model, we trade model complexity and the known complexity of brain networks for a straight-forward interpretation of the model's output. More generally, we might consider a family of nested models, each nested model being strictly more simple than its parent, and perform model selection across this space to determine which model is best supported by the available data [49]. However, whatever the resulting model that is most supported, we must still subject this model to out of run prediction. Otherwise, we will only have determined the model that best fits the training data rather than a measure of how well the model is capturing fundamental network structure.

Dynamic causal modeling is one particular framework for mathematically understanding brain activity. DCM is appealing because it combines both a graph representation of the brain with a compact representation of the workings of each node using the bilinear differential equation. However, DCM still has its difficulties (sufficient data support for optimization, the challenge of infrequently sampled fMRI data, and the selection of timeseries from regions of interest). From the perspective of developing biomarkers or insight into speech models from which biomarkers may be derived, it is not the only option.

We do believe that any technique preserve network (graph) representation of the brain. We also believe in the validation of any technique by out of run prediction of the BOLD timeseries as well as using model parameters as classification features. Within these constraints, we propose using other data collected from the subject to constrain the network graph. For example, nodes and input time series could be defined from an analysis of connectivity in resting state scan or from a different speech task by the subject. Furthermore, the idea of a single spatial scale for representation is limiting. Rather than arbitrarily choosing a set of clusters for the graph, it may be helpful to approach the parcellation as a hierarchical clustering problem in which the hierarchy with multiple spatial scales is preserved. This idea follows from viewing brain dynamics as being generated from a simpler underlying structure, just as DCM does, but as part of discovery of the underlying structure, the multi-scale nature of the data is taken into account. Seversky et al. [117] and others explore these ideas as part of topological data analysis and time series embedding techniques.

Chapter 4

Phoneme Rate Model

THIS chapter explores a neurocomputational model for the phoneme rate biomarker. We propose interpreting phoneme rate control within an existing neurocomputational model of speech production, the Directions into Velocities of Articulators (DIVA) model. Furthermore, we extend the DIVA modeling framework with an algorithm to estimate subject-specific model parameters for rate control. We use all estimated model parameters as unobserved or latent biomarkers for assessing depression severity.

Section 4.1 discusses background definitions and current models of speech production. Section 4.2 describes a conceptual model of phoneme rate control. Section 4.3 provides an algorithm for inference of the latent rate parameters from a neurocomputational model of speech production. Sections 4.4 and 4.5 present results and a discussion of the results from experiments modeling the phoneme rates of depressed and control subjects.

Our aim was to use concepts of phoneme rate modeling to derive unobserved biomarkers of depression. Our contributions include a quantitative algorithm for doing so, and more generally a discussion on the limitations and extensions of our specific model. From these results, we are led to alternative means of investigation and modeling of rate control in speech.

■ 4.1 Background and Prior Art

There are three neurocomputational models of speech production that are relevant to this thesis. They are the Directions into Velocities of Articulators (DIVA) model version 1.0 [59] and DIVA 2.0 [61], and the Gradient Ordered DIVA (GODIVA) model [16]. To understand their function, we will review basic speech production concepts in Section 4.1.1. Then, in Section 4.1.2 we will summarize the contributions and limitations of these models to highlight our own contribution to modeling the phoneme rate biomarker.

■ 4.1.1 Definitions

The vocal tract, comprising the airway above the vocal folds, the lips, jaw, tongue, and velum, can be approximated as a series of connected tubes of different diameters [28]. Just as a glass filled with different levels of water has different resonant frequencies, so too does the vocal tract have different resonances depending on its shape. The

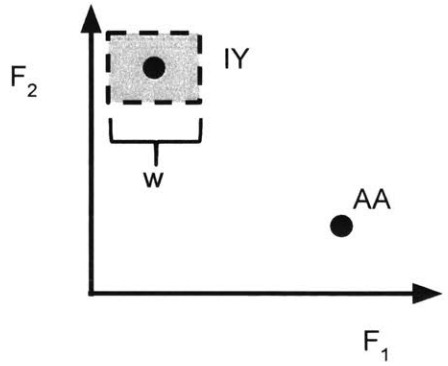


Figure 4.1: An example of two phonemes in F_1 and F_2 space. The auditory target region (dotted grayed box), and the auditory width parameter, w, for the F_1 dimension are shown for the [IY] phoneme.

resonances of any particular vocal tract configuration are called formants. The lowest frequency resonance is the first formant F_1, the next highest resonance is F_2, and so on.

There is a tight mapping between the values of F_1 and F_2 and the percept of different phonemes. The vowels of English speech can be represented as regions in a graph with x-axis, F_1, and y-axis, F_2 as in Figure 4.1. Consonants can also be placed in a F_1-F_2 coordinate system but the mapping is more variable and context dependent than the mapping for vowels.

The auditory width of the phoneme target is the size of the region of auditory space in which one sound is classified as one phoneme vs another. We use the term "target" to suggest that a goal of speech production is to place a desired sound at a particular point in auditory space. While a region of acceptability defined by the width surrounds the point target, we adopt the view that the target itself is a point. This view is supported by the work of Niziolek et al. [96] who found that even within the production of a vowel phoneme, speakers alter their production to bring the F_1 and F_2 formants closer to a canonical point in acoustic space.

Speakers may have several goals during communication such as a somatosensory goal in which a sound should have a particular tactile feel (e.g., the tongue hitting the back of the teeth such as when pronouncing "tea"). In the general case, the acoustic and somatosensory targets may be a dynamic trajectory that is a function of time rather than a fixed point. An example of a dynamic trajectory for a phoneme is the [W] in "way" ([W] [EY]) [60, 81]. An acoustic program is the timeseries of individual acoustic point targets that are intended to be executed, and similarly, an articulatory program is the timeseries of target positions for each of the articulators. Without loss of generality, we focus on auditory targets as the dominant goal in modeling low level speech production.

Because phonemes are so elemental to production, and some words occur so often, the concept of a motor chunk has been hypothesized. A motor chunk is an overlearned articulatory and acoustic program that can be as short as a phoneme or as long as a polysyllabic sound. A motor chunk is synonymous to a speech unit within the DIVA literature [60].

To this point in the thesis, we have focused on individual phonemes. However, words and sentences are a continuous stream of phonemes. Speech sequencing is the process by which motor chunks are produced one after the other [60].

A model parameter is a number or function that controls the relationship between the model's inputs and outputs. For example, the auditory width of an acoustic target is a parameter for the component of a model that monitors auditory error. The component takes as input the error itself (the difference between the perceived acoustic production and the auditory target) and outputs a modified error signal that depends in part on the auditory width parameter. For example, the output might be a binary decision about whether an error was registered or not. An example of a function parameter would be the specification of a means for transforming a neural command into an acoustic production. The functional form might be a multivariate linear equation, or a physics based simulation.

Speech production occurs at multiple levels that can be described on a continuum from high level to low level control pictured in Figure 4.2. High level control is control of timing at the longest timescales such as over the course of reading an entire passage or giving a presentation. At the next lower scale is control of timing within a sentence. At the next lower scale is control of timing within a word. Word level control in particular depends upon properly placing phonemes in a sequence. At the lowest level of control is the phoneme.

The idea of a motor chunk is connected to the idea of these time scales. A motor chunk is generally believed to be at the word or phoneme level of control. In this thesis, we will use low level control to refer to the motor chunk timescale, and high level control will refer to sequences of motor chunks. Generally, this can be thought of as low level control at the phoneme level and high level control at the sequence of phoneme level, but in some cases the motor chunk may be larger than a single phoneme[1].

■ 4.1.2 Current Neurocomputational Speech Production Models

DIVA 1.0 [59] and DIVA 2.0 [61] are neurocomputational models of low level speech motor control. They are defined at the motor chunk level and describe the production of a speech motor chunk. GODIVA is a neurocomputational model of high level motor control. GODIVA is defined at the sequence of motor chunk level and describes how a sequence of motor chunks may be produced. Specifically, in GODIVA, a sequence of motor chunks is retrieved from long term memory, and each individual motor chunk

[1]Timescales may be further divided into groupings of syllables or phonemes in, for example, consonant vowel consonant groups and other representation such as onset and rhyme which are outside the scope of this thesis [60].

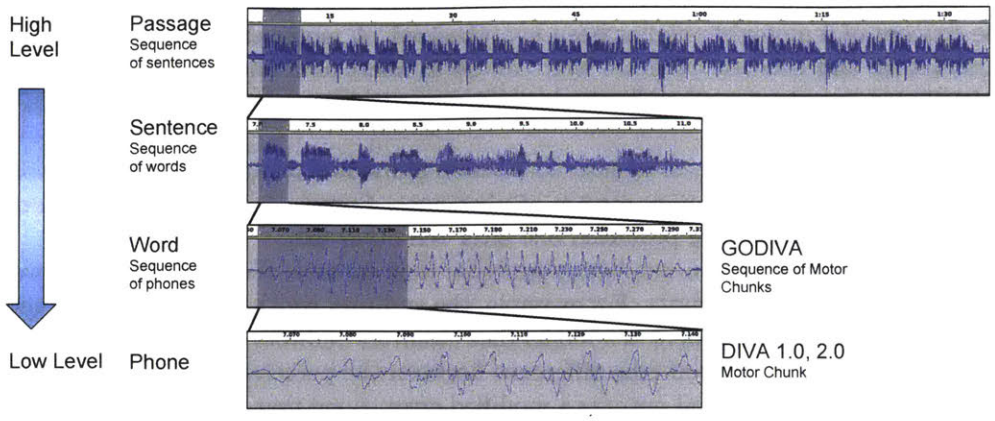


Figure 4.2: Speech control occurs at multiple time scales from high level, paragraph timescales to low level, phoneme timescales. DIVA operates at low level control, and GODIVA operates at the next higher level from DIVA.

is held in working memory simultaneously (aka in parallel). GODIVA provides an explanation for how a parallel representation of speech chunks can be serially produced [16]. DIVA 1.0, 2.0, and GODIVA have both been developed with the aid of human functional magnetic resonance imaging experiments in the sense that components within the models have been localized to different cortical and subcortical structures. We summarize and expand on the attributes of these models listed in Table 4.1.

DIVA 1.0 has important notions of rate and touches upon how these ideas related to sequences but DIVA 1.0 does not have a method for estimating subject specific parameters from speech. In DIVA 2.0, the idea of a sequence of motor chunks is not present. DIVA 2.0 is only concerned with the method of execution of an individual, specific motor chunk. To GODIVA, the idea of a motor chunk exists as an abstraction. Motor chunks are used as undefined placeholders that GODIVA integrates into production templates to produce in a serial order.

DIVA 1.0 and 2.0 both touch upon overall speech rate, and DIVA 1.0 in particular explains speech rate with an additional level of detail. DIVA 2.0 is the simplest to understand because DIVA 2.0 is a pure mimic of a given input speech signal. Consequently, DIVA 2.0 has no notion of rate or how variability of rate might occur. DIVA 2.0 learns to duplicate the acoustic characteristics of the template signal. The template signal is whole-sale learned as one speech chunk, even if that speech chunk is unrealistically long in duration (e.g., a paragraph). Nonetheless, once DIVA 2.0 has learned the speech chunk, it does have rudimentary rate control in the form of an overall read-out speed of its learned motor program. The GO signal is the overall readout speed, which we will represent with α. By analogy, DIVA 2.0 is like a record player that can speed up or slow down a song uniformly [59], but at present does not have an internal control

Table 4.1: Comparison of neurocomputational speech production models

Attribute	DIVA 1.0 [59]	DIVA 2.0 [61]	GODIVA [16]
Sequence	Partial	No	Yes
Motor Chunk	Yes	Yes	No
Rate	Yes	Partial	No
Method for Setting Model Parameters	No	No	No
Account for Affect or Depression	No	No	No

mechanism to adjust rate. As shown by Cai et al. [22], humans do have the capability of making syllable level adjustments in rate.

Mathematically, let the amplitude vs time speech waveform be represented as $x\left(\frac{t}{\alpha}\right)$. Then the GO signal is a time compression (increase in speaking rate) or time dilation (decrease in speaking rate) depending on whether α is greater than or less than one. Because α causes a uniform change to the whole signal x, all measures of speaking rate whether words per second, syllables per second, or phonemes per second, are all increased or decreased by the same proportion dictated by α. The velocity strategy for speaking rate control is the name for the method of changing speech rate by changing α [59].

Human speakers will move their articulators faster to speak faster (i.e. achieve auditory targets faster) [59], so the velocity strategy should be a part of a model of speech rate. However, there is substantial evidence from literature that humans do not articulate in the same manner at different speaking rates as DIVA 2.0 would do, and there is evidence that humans do not uniformly speed up or slow down all phonemes equally within an utterance, as DIVA 2.0 would do [59]. Therefore the velocity strategy is not a complete explanation of speaking rate control.

In addition to describing the velocity strategy, Guenther [59] described a second strategy used by human speakers in fast speech, and that is the amplitude strategy. The amplitude strategy is the change in the maximum excursion of the articulators from a neutral position to change speaking rate, and the amplitude strategy can be modeled using the auditory width parameter. A larger auditory width corresponds to a smaller articulator travel distance before the target is achieved and the next target can be produced. Therefore, a large auditory width allows achieving more auditory targets per unit of time than a small auditory width, and consequently results in a faster speaking rate [59].

Consonants and vowels have auditory target regions that significantly differ in their shape as pictured in Figure 4.3. The difference in shape gives rise to different instead

of uniform changes in phoneme durations when speaking rate increases or decreases. Consonants have narrow auditory targets that are inflexible to significant shrinkage or expansion along certain dimensions due to the precise timing required to execute a consonant properly. In other words, a consonant may still be effectively produced along a range of possible F_1 values, but the F_2 value is inflexible to much change. By contrast, a vowel auditory target region can be relatively expanded or contracted in all directions without significantly compromising the sound [59].

As a consequence of the amplitude strategy and the differences in intractability of consonants vs vowels, DIVA predicts differential changes in duration for different phonemes for changes in speaking rate. When a person is speaking quickly, the person might expand the auditory target regions for consonants and vowels, but this has the effect of expanding the region more around the vowel than the consonant. Consequently, the distance an articulator must move through auditory space from a starting point to the boundary of the vowel or consonant target will be differentially changed for consonants vs vowels. The distance from a starting acoustic position to the boundary of an enlarged vowel target will be shortened relatively more than the distance between a starting position and an enlarged consonant target. Therefore, vowel targets will be achieved more quickly compared to consonants. The end result will be greater contraction of vowel durations than consonants during fast speech. The velocity strategy does not account for differential changes in phoneme duration because the velocity strategy predicts that all phoneme durations are changed uniformly [59].

GODIVA, because it is a sequence model, does not concern itself with the rate of its individual motor chunks. In GODIVA, a chunk is cued for production, and DIVA executes the chunk at a rate dictated by the chunk itself and DIVA's α parameter [16].

■ 4.1.3 Innovation and Contributions

We propose to model phoneme rate within the context of DIVA in order to use model parameters α and w as features for assessing depression. We propose to use α and w within DIVA as a means to derive phoneme dependent changes in rate. The phoneme rate effects occur through the velocity strategy and amplitude strategy proposed by Guenther [59]. Our model is successful if we can find α and w that are stable across speech productions for an individual as we would expect these parameters to be relatively constant for a person. Furthermore, our model is successful if the mean absolute error between predicted and measured phoneme rates for a production is less the mean absolute error computed using a person's average phoneme rate. In other words, our model should capture phoneme specific variation. To date, phoneme specific variation has not been shown to be captured by the DIVA model, nor has this mechanism of quantitatively capturing phoneme specific variation been proposed or assessed.

If our model is able to capture phoneme specific variation that is stable within an individual, we will assess the relationship of these model parameters with respect to the person's depression severity. To confirm our hypothesis that the model parameters are capturing an aspect of mental health, we need to be able to predict depression severity

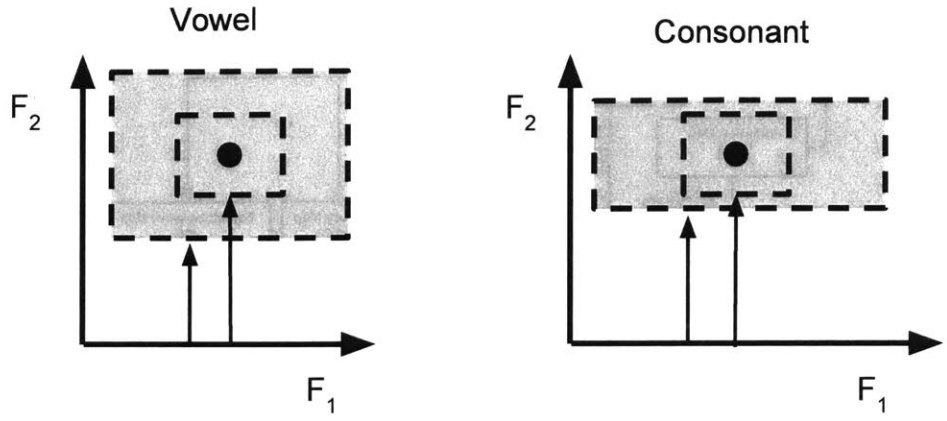


Figure 4.3: An initial and expanded auditory target region for a vowel (left) and consonant (right) in F_1 and F_2 space. The difference in regions is a result of differences in production requirements for vowels vs consonants. An expanded vowel target causes a greater reduction in travel distance (arrow length) from a starting point with a smaller F_2 than the target relative to an expanded consonant target. Consequently vowels are shortened in duration more than consonant in fast speech. After [59].

from the parameters (e.g. classify depressed vs non depressed subjects or correlate BDI with a function of the model's parameters). Of course, it may be overly optimistic to suggest that depression can be fully characterized by two model parameters. Realistically, if phoneme rate control was a clear biomarker of depression, then we would expect these parameters to feature prominently in any prediction system that drew upon an array of features.

To our first contribution of modeling phoneme rate specific effects, we add a second contribution: an algorithm for estimating model parameters from a single speech utterance. The DIVA models currently have no way of simply estimating either α or w. For a system to practically use our model-based features, we need to be able to derive these features without independent experiments on the subject. We need to be able to derive the features from the speech sample we have from the subject. We propose a novel algorithm for deriving these parameters.

■ 4.2 Models of Phoneme Rate

We propose a neurocomputational model of phoneme rate variability that depends upon low level speech motor control. We call this model the "execution model" to emphasize the importance of low level motor control mechanisms within DIVA in contrast to higher levels of control in GODIVA. While GODIVA also may be used to model phoneme rate,

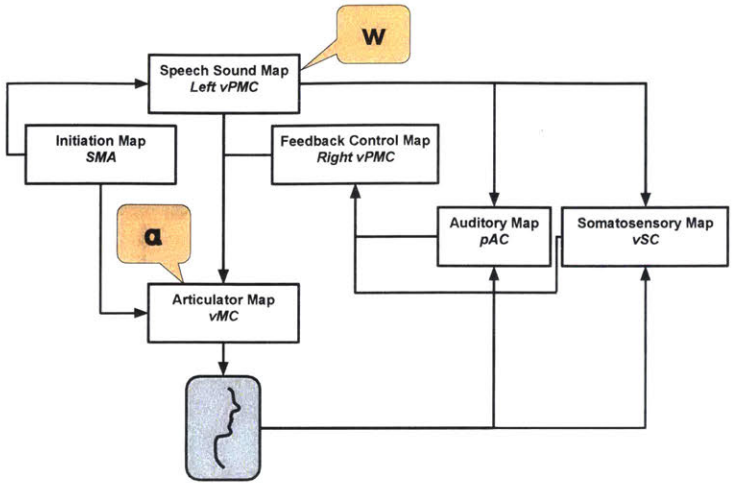


Figure 4.4: Schematic version of the DIVA model after [60] with the hypothesized locations of the two neural phoneme rate control variables, w and α.

we chose to focus on the DIVA model. This decision was in part because of preliminary (and then unreproduced) evidence of phoneme rate control at the cortical level in our subjects from the fMRI experiments and because we could leverage our prior experience working with the DIVA model.

■ 4.2.1 The Execution Model

The execution model explains phoneme rate changes using a combination of the amplitude and velocity strategies covered in the background. We associated with the velocity strategy a parameter that is the GO signal strength, α. We associate with the amplitude strategy a measure of the auditory target width, w. We summarize the model in Tables 4.2 and 4.3 and describe it further below.

Figure 4.4 shows the potential locations of α and w in the DIVA 2.0 model. w naturally affects auditory processing, so while evidence of it may appear in auditory cortex, we actually would localize it to premotor cortex. In the DIVA model, the premotor cortex sends auditory expectations to auditory cortex, but w likely also influences articulation to some extent. Premotor cortex sends projections to both auditory cortex and motor cortex, so locating w in premotor cortex allows w to influence both systems [128]. α is not explicitly located in the brain by Tourville and Guenther [128], but because α is part of general motor drive, α likely would correlate with amount of neural activation within ventral motor cortex even if α is not controlled within the ventral motor cortex.

We can explain phoneme rate decrease by decreases in α or by increases in w. When

interpreting the effect of w on phoneme rate, we mention Fitts's law [45]. Fitt's law is a widely replicated law of human motion, however, it holds only imperfectly in speech production [77]. We state a linear formulation of Fitts's law below:

$$\text{Movement time} = a\,(\text{Index of Difficulty}) + b \tag{4.1}$$

We also define the Index of Difficulty (ID) in terms of the distance to the target, D, and the width of the target w using MacKenzie [85]'s formulation:

$$\text{ID} = \log_2\left(\frac{D}{w} + 1\right). \tag{4.2}$$

Under Fitts's law with constant distance between targets, an increase in w implies a decrease in ID, which in turn leads to a decrease in movement time (MT). This is consistent with the idea of "hitting" the boundary of an auditory target sooner and therefore being able to progress to the next auditory target.

We can use α and w in the execution model to explain two prominent speech changes in depression: a general slowing of production [94], and a general loss of articulatory precision [114]. The precision of a motor target is inversely proportional to the width of the auditory target region. From any given starting position, movement into the goal region requires less of an excursion for execution of the phoneme to be considered complete when the width is large. The overall loss of precision can be quantified by the area in acoustic space encompassed by vowels that have extreme acoustic target points (e.g. [IY, AA, UW]). The area of the vowel space triangle can be computed using the length of the triangle's three sides (a, b, c) and Heron's formula: $A = \sqrt{s(s-a)(s-b)(s-c)}$, $s = \frac{a+b+c}{2}$ [114].

While there is a general slowing of phonemes in depression, not all phonemes are lengthened in duration. Trevino et al. [129] reported that [AA], when significantly correlated with the total HAMD score, decreased in duration (but see our results in Chapter 2 in which we found AA increased in duration). Future work will need to provide an account of how some phonemes can be shortened and others lengthened. A two parameter execution model is not flexible enough to account for this observation.

To explain the decrease in phoneme rate and the loss of articulatory precision in depression, we hypothesize a decrease in α and an increase in w, but with the decrease in α dominating the speed increase that comes from an increase in w. Literature suggests these are possible explanations consistent with depressed behavior. Buyukdura et al. [21] summarizes in a review that psychomotor retardation is a consistent indicator of depression where psychomotor retardation is the general slowing of motion and also thought. Psychomotor retardation corresponds to a decrease in α. Schroder et al. [115] and Bailey et al. [8] both found evidence of reduced error awareness in subjects with depression. Specifically, Schroder et al. [115] used an electroencephalogram study to show that task difficulty and depressed state resulted in lower error related positivity signals (Pe) which are electroencephalography event related potentials that occur upon error commission. Bailey et al. [8] found a similar result of reduced Pe signals in subjects

Table 4.2: The execution model as a potential model of phoneme rate control.

	Execution Model
Predict phoneme duration is context dependent	Yes
Neurocomputational framework	DIVA
Neural basis	Cortical (pre-motor cortex, and ventral motor cortex)
Mechanism for changing phoneme duration:	Change in α and/or w

with traumatic brain injury and depression but not with depression only. Reduced error awareness in depression corresponds to an increase in w.

It should be noted that there are two variants to the execution hypothesis. Under variant one, as we have described, the auditory goal regions are modulated which produce the emergent effect of imprecision. However, under variant two, the auditory goal regions are actually still the same as in normal controls. Instead, the cost of reaching into the auditory regions is greater than the cost of registering an error. Therefore, even though an auditory error is measured during feedback, this is preferred over the cost required to eliminate the error. A mismatch study in which a subject's speech is perturbed and error responses is recorded might disambiguate between these two hypotheses as might a perceptual acuity study in which depressed subjects have to classify presented sounds that span a region in auditory space. We define perceptual acuity as the ability to accurately classify sounds along a continuum in auditory space.

■ 4.3 Methods

We turn in this section to implementing the execution model in order to quantitatively estimate α and w from subject speech. We surmise that by modeling subject specific phoneme variability, these two parameters may be able to separate depressed from control subjects as schematically shown in Figure 4.5.

Reducing the complexity of speech articulation differences between depressed and control individuals to a lower dimensional space comprising these two parameters is the ultimate goal. However, as a reasonable condition for using these parameters, we require that the estimated model parameters do replicate subject specific phoneme variability. We estimate α and w by minimizing the error between predicted phoneme durations, $\hat{\boldsymbol{d}}$, and the measured phoneme durations, $\boldsymbol{d}$, for the phoneme acoustic targets, $\boldsymbol{p}$, in the speech to be modeled. Our computational model that generates $\hat{\boldsymbol{d}}$ from α, w, and $\boldsymbol{p}$ is $f(\alpha, w; \boldsymbol{p})$. Formally, our optimization objective is

Table 4.3: Parameters in the execution model and their effect on phoneme duration.

α	w	Phoneme Duration	Description	Consequence
$\uparrow$	-	$\downarrow$	Articulators travel faster	$\uparrow$ formant velocities and accelerations
-	$\uparrow$	$\downarrow$	Smaller travel distance	Poor perceptual acuity. Vowel space reduction.
$\downarrow\downarrow$	$\uparrow$	$\uparrow$	Depression causes a large decrease in α that dominates speed increase from $\uparrow$w. $\uparrow$w still causes loss of acuity and vowel space reduction.	

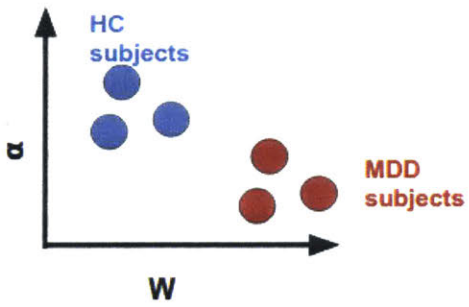


Figure 4.5: Hypothesized separation of subjects in latent parameter space. Controls (HC) have narrow auditory targets (small w) and an agile, responsive motor system (large α) in contrast to depressed subjects (MDD).

$$\alpha^*, w^* = \text{argmin}_{\alpha,w}\text{Error}\left(f\left(\alpha, w; \boldsymbol{p}\right), \boldsymbol{d}\right). \tag{4.3}$$

We use lower case, non-bold letters (e.g. α, f) to represent scalars or functions, and lower case, bold letters to represent vectors (e.g. $\boldsymbol{d}$).

The entire estimation process is shown schematically in Figure 4.6, and we use the following sections to describe each part of the estimation process. Section 4.3.1 describes estimation of $\boldsymbol{d}$, and Section 4.3.2 describes creation of the phoneme acoustic targets $\boldsymbol{p}$. Section 4.3.3 describes the computational model, f, in detail, and Section 4.3.4 describes our solution to the optimization objective in equation 4.3.

■ 4.3.1 The True Phoneme Durations

The algorithm begins with an acoustic production from a subject with unknown depression severity. We derive from this utterance the true phoneme durations, $\boldsymbol{d}$, that we will model with f. We use the automatic, transcript-free phoneme recognizer [118] from Chapter 2 to identify the phonemes within the utterance and to estimate the duration, d_i, of each phoneme in the utterance. We collect the true durations into an ordered vector $\boldsymbol{d}$. We emphasize "ordered" because we preserve the order of the phonemes. We do not compute a mean phoneme duration if the phoneme occurs more than once in the utterance as we did in Chapter 2 to derive a phoneme rate feature. In other words, $\boldsymbol{d}$ is N phonemes long where N is the number of phonemes in the utterance rather than N always equal to 40 (39 phonemes plus silence) as in our feature vector in Chapter 2.

■ 4.3.2 The Nominal Acoustic Trajectory

The nominal acoustic trajectory, $\boldsymbol{p}$, is a sequence of target formant values. Each element, p_i, in the sequence is a (F_1, F_2) pair that corresponds to the (F_1, F_2) for the i^{th} phoneme identified in the utterance. Importantly, the target (F_1, F_2) for a phoneme is not derived from the subject's utterance. Instead, (F_1, F_2) is the average (F_1, F_2) for all instances of that phoneme identified in all healthy control subjects from the Rainbow passage.

We acknowledge that by using an average formant value across a diversity of subjects (i.e., not accounting for vocal tract differences, or individual variability even among controls), we are introducing error into the model. However, the DIVA model uses a particular vocal tract, and we would like to see acoustic targets represented within the parameters of that vocal tract space.

We derive the average phoneme target only from control subjects instead of depressed subjects to avoid a bias of the formant targets from depressed subjects. Vowel space reduction is a noted phenomenon in depression meaning that a given vowel generally has more centralized formant values for a depressed subject than a control subject [114]. Vowel centralization is the clustering of formant values for all phonemes towards a central cluster in formant space. Assuming that depressed and control subjects have the same point target, an estimate of that point target from depressed subjects would

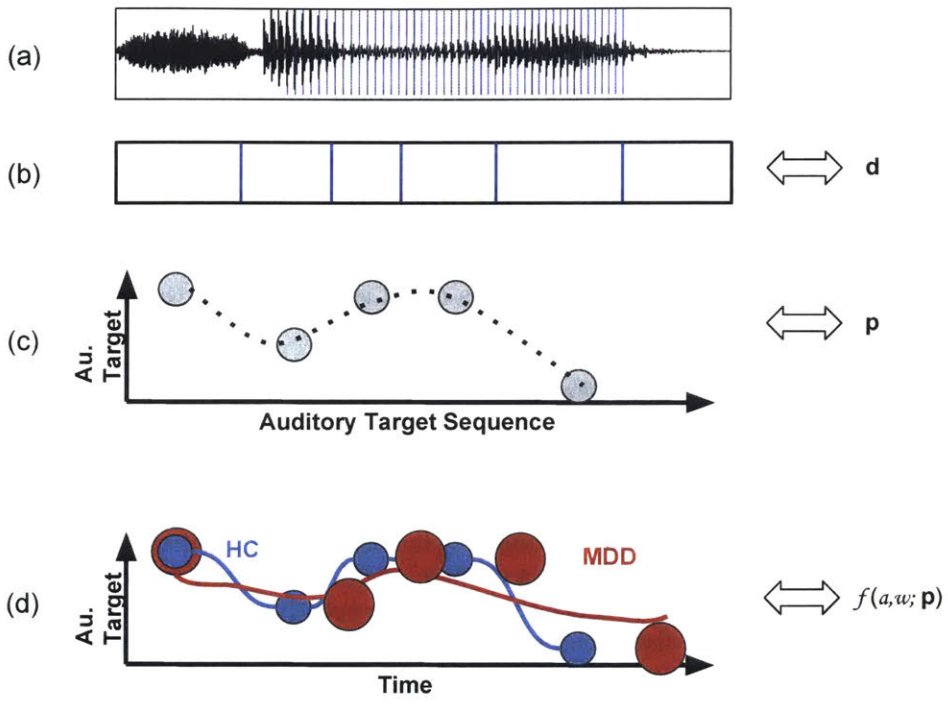


Figure 4.6: Algorithm for estimating w and α illustrated schematically with hypothetical data. (a) Input waveform from the subject. (b) Automatic phoneme recognition. (c) Sequence of phonemes and their corresponding auditory targets. phoneme duration is not specified, only phoneme order. (d) Output from optimization of the model for a control vs depressed subject. The depressed subject has large auditory targets that are only imprecisely attained as opposed to the narrow, precisely met targets of a healthy control. Because of differences in the latent parameters, the underlying model yields differences in phoneme duration. Consequently the durations for the depressed subject are increased and the time series appears stretched relative to the control.

be biased towards a centralized point rather than the true target.

On a practical level, Hillenbrand et al. [67] has shown considerable variability in production of vowels in auditory target space among sexes. We used a common estimate of the formants among controls as the target because of limited data. A form of vowel space normalization may be warranted as such. Another point is that the DIVA articulatory speech synthesizer models the articulator to acoustic mapping for one set of articulators- the set for which the model was created. Individual speakers will have different vocal tract properties that can alter this mapping, and consequently alter what the "optimal" path in articulator space should be when traversing acoustic space. To see whether performance could be improved, a vocal tract model that allowed subject specific anatomy could be used such as the one by Birkholz [12].

A final word about the nominal acoustic trajectory is that duration is not part of the definition of $\boldsymbol{p}$, only the order of the phonemes. The point of optimizing the computational model is to find α and w such that using only the order of the phonemes and their acoustic targets the predicted durations $\hat{\boldsymbol{d}}$ will agree well with the true phoneme durations $\boldsymbol{d}$.

■ 4.3.3 The Computational Model

The purpose of the computational model, $f(\alpha, w; \boldsymbol{p})$, is to convert a given α, w, and nominal acoustic trajectory $\boldsymbol{p}$ into a set of predicted phoneme durations $\hat{\boldsymbol{d}}$. We base our model upon the DIVA architecture, but we implement several modifications to adapt the existing DIVA 2.0 architecture to support phoneme rate generation.

The current DIVA architecture is a "mimic" model. Given a speech utterance, the current DIVA model will produce that utterance exactly as one continuous target without distinguishing the individual phoneme units from one another. In our model, we will use the basic architecture of DIVA 2.0 but incorporate the ideas of the velocity strategy and amplitude strategy discussed earlier through the α and w parameters. Furthermore, we will use the idea of discretely changing phoneme targets where each discrete target represents a phoneme instead of the current method which uses a fully specified set of formant tracks.

We describe the computational model by beginning with the basic architecture. Then, we describe our modification to support switched phoneme targets and the incorporation of the w parameter. After that, we describe how we incorporate the α parameter. We finish by describing the vocal tract model.

The Basic Smith Predictor

We model phoneme rate variability using a feedback control system as shown in Figure 4.7 after [43, 60]. A feedback control system is the collection of a reference signal, r, a controller c, and a plant g linked in such a way that the plant's actual output, y_p, or estimated output, y_{fb} is available with the reference signal to the controller. The objective of feedback control system is to make the produced output match the reference signal. The controller is a function that takes as input the reference signal

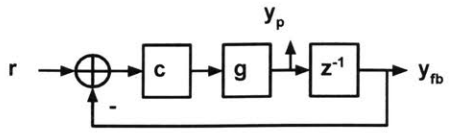


Figure 4.7: Simple feedback control system with controller, c, plant, g, reference, r, plant output, y_p, and feedback signal, y_{fb}.

and a feedback signal and produces an output signal that goes into the plant g. The plant g is a function that takes as input the control signal and outputs some quantity y_p[2].

The objective of the controller is to facilitate the alignment of the produced and desired output in such a way that the entire system has other desirable properties (e.g., in steady state, $y_p = r$, or perhaps a steady state error is permitted in exchange for a quick approximate match of r to y_p).

In our model, for each individual phoneme, p_i, our model has the reference acoustic target, $r = (F_1, F_2)$, and an estimate of its current position, y_{fb}, in acoustic space. Our plant, g, is a vocal tract model that will convert articulator positions into formant frequencies and is discussed in detail in a later section. The controller, c is the subject of this section and the following two sections.

The simple feedback control system in Figure 4.7 essentially has no delay between the output of the plant and the feedback of that output to the controller. There is a one sample delay represented in discrete time as z^{-1} in the diagram since the simulation is implemented as a discrete time simulation in software. If there were a large delay, intuitively, this could be a serious problem. Imagine if there were a large delay between adjusting the hot and cold knobs of a shower. Achieving the desired temperature must be done slowly in order to give the new knob setting time to affect the water temperature.

In humans, though our auditory perception is excellent, there actually is an appreciable delay between when we say something and when we hear ourselves say what we said. In order to handle this delay, our nervous system uses an internal model of the plant to predict the output that we think we will actually generate if we use a control

[2]True speech production has both feedforward and feedback components Guenther [60]. Feedforward control is the use of a preconceived, already learned control signal as an input to the controller. In pure feedforward control, the plant's output is changed independent of whether the output matches the reference signal. The plant simply obeys the current signal it receives from the controller. However, we focus on feedback control as an explanatory mechanism for building in the flexibility of generating phoneme rate variability because our premise is to model phoneme production when the controller's feedforward plan is not known. If we had the feedforward control signal, there would not be anything to model regarding phoneme rate variability. The feedforward signal must be learned, and feedback control is a method for learning.

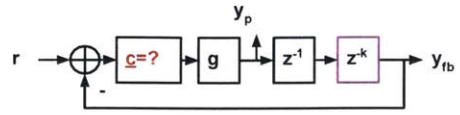


Figure 4.8: Smith's problem: c was designed for a system without a bulk delay. How should the new c=? be chosen such that the system with the bulk delay behaves as if c were the controller for a system without the bulk delay?

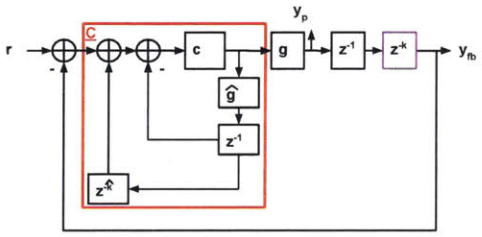


Figure 4.9: Smith's solution: the original c can be reused in the system with a bulk delay by using a prediction model of the plant, $\hat{g}$, and an estimate of the bulk delay, $\hat{k}$,

signal. This expected feedback is then combined with actual feedback when it arrives to create a stable but responsive control signal [53, 103, 146].

If a system designer had designed a controller, c, that met their design specifications, and then tried to use that controller in Figure 4.8, the result would not be as intended because of the bulk delay z^k. Smith [121] invented a solution to this problem pictured in Figure 4.9 [136] that uses the original controller c, a predictive model $\hat{g}$, and an estimate of the bulk delay $z^{\hat{k}}$. Using Smith's architecture, the controller $\underline{c}$, acts in the presence of the bulk delay like c acts when no bulk delay is present. A crucial element of the Smith architecture is an accurate predictive model, $\hat{g}$, and an accurate estimate of the true bulk feedback delay $z^{\hat{k}}$. In our simulation, we assume these values are known, but it is an area of open research of how best to estimate these values.

The Smith Predictor with Changing Targets

Thus far, we have introduced a feedback controller that will adapt the articulator positions to match the acoustic target of a single phoneme. We need to modify this architecture to account for the sequence of phonemes the model must generate, and in our modification, we will introduce the auditory width w.

Our modification is a straightforward implementation of the idea of an acceptable acoustic goal region around the auditory point target. Once the estimated feedback y_{fb}

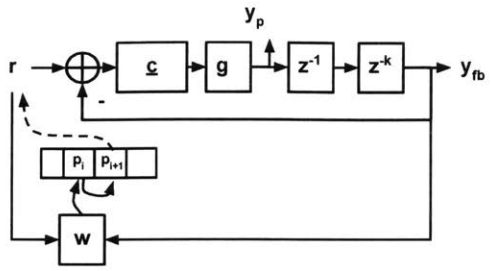


Figure 4.10: The Smith predictor updated to use the auditory width parameter, w, inside a comparator block, **w**, that controls the current target phoneme.

is within w of the target, the goal is considered achieved Guenther [59]. The achievement of the goal sends a signal to a buffer that holds the phoneme targets $\boldsymbol{p}$, and the next phoneme target p_{i+1} becomes the new phoneme target of the feedback control system. Figure 4.10 shows this modification. The duration of time from the achievement of the previous phoneme to the completion of the current target phoneme is the duration of the current phoneme target.

To use w, we need to quantitatively define the acoustic error expression

$$e_{Au} = \text{Error}(r, y_{fb}) \tag{4.4}$$

which captures the difference between the reference and the estimated positions. As a first attempt, we used the Euclidean distance between r and y_{fb} but we quickly discovered a problem due to the difference in span of the vowel space along the F_1 and F_2 axes. In acoustic space, vowels only vary by several hundred Hertz along the first formant dimension but they can vary by over a kilohertz in the second formant dimension. A Euclidean distance was placing a circle of acceptability around the target, so phonemes that are actually quite distinct along the F_1 dimension were being considered equivalent.

To counter this effect, we use a weighted distance metric that accounts for the differences in scale along the two formant axes:

$$e_{Au} = (\mathbf{r} - \mathbf{y}_{fb})^T \boldsymbol{\Sigma} (\mathbf{r} - \mathbf{y}_{fb}) \tag{4.5}$$

where Σ is a diagonal matrix with main diagonal elements of 1 and 1/9 which approximate the differences in vowel spread along the two formant axes. The values were chosen by looking at a scatter plot of vowel points from the subjects. If $e_{Au} \leq w$, then the phoneme target is considered achieved, and the next phoneme in the buffer is used to update r.

We note that reducing a phoneme's complete description to two formant values loses considerable spectral detail and consequently detail about the phoneme's manner of articulation. However, this simplification is based on established mapping of vowels

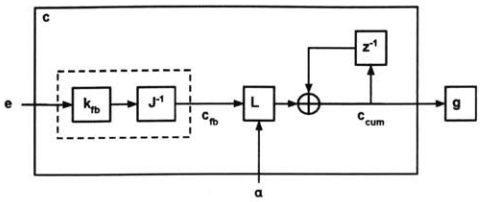


Figure 4.11: Inside the controller, c, which features a proportional error mechanism (dotted line), an inverse Jacobian for converting between errors in auditory space to updates in articulator position space, the limiter, L with parameter α, and an accumulator that together make c act as a proportional-integral controller.

in particular to a two-formant axes, and provides a tractable starting point for investigating timing control principles. An extension of this two parameter feature vector for a phoneme to a larger feature vector can still be performed within this framework.

The Controller in Depth

Up to now, we have treated the controller, c, as a black box. We now go inside the controller and describe how it functions and how it incorporates the GO signal strength α. Recall that the purpose of the controller is to change the articulatory parameters of the plant g such that the output of the plant matches the reference target. DIVA's feedback controller is a proportional error feedback controller. In other words, the change in the magnitude of the control signal is directly proportional to the magnitude of the error signal. The larger the error signal, the larger the change in the control signal.

Because the articulators in DIVA maintain their current configuration instead of relaxing to a neutral configuration, DIVA actually acts as a proportional plus integral controller. This kind of controller, as opposed to proportional alone, is what allows DIVA to achieve zero steady state error for a constant target which is what we have for each phoneme.

Figure 4.11 shows the inside of the controller. The dotted line surrounds the proportionality constant k_{fb} that is applied to the error signal and a second function, an inverse Jacobian J^{-1} that converts between auditory space and articulator position space. At the most basic level, the inverse Jacobian "makes the units match" by converting from units of Hertz for the error signal to units of articulator position for the control signal.

Formally, the Jacobian is a matrix of gradients of the acoustic signal with respect to the control parameters:

$$J = \begin{pmatrix} \frac{\partial F_1}{\partial x_1} & \cdots & \frac{\partial F_1}{\partial x_7} \\ \frac{\partial F_2}{\partial x_1} & \cdots & \frac{\partial F_2}{\partial x_7} \end{pmatrix}. \tag{4.6}$$

The first row of equation 4.6 is the derivatives of the first formant with respect to each of the control parameters in the plant. The second row is the derivative of the second formant with respect to each of the control parameters. In words, the Jacobian describes how changing control parameters will change the output of the plant. However, the controller needs to perform the inverse operation. The controller has a desired change in the outputs of the plant and needs to determine the corresponding change in the control parameters. Therefore, an inverse Jacobian is used. Because the Jacobian is not square in our case, an inverse does not exist, so we use a numerical approximation to the inverse Jacobian.

Figure 4.11 also shows an accumulator that tracks the current state of the articulators. The accumulator mathematically keeps the articulators in their current position until the controller gives a new control signal. It is implemented as a discrete running sum of all past control signals.

The final element of Figure 4.11 is a limiter, L. The limiter is the key element that incorporates the α parameter as well as making possible variability in phoneme rate for different phonemes depending on the relative position of the phonemes in auditory space and w. In a proportional error feedback control without a limiter, no matter how large or how small the error, the control signal will be scaled as determined by the scaling constant $k_{fb} < 1$. However, this is inappropriate for our model for two reasons. First, it is unreasonable that the physiological system has no bound on how much the articulators can move in a given unit of time. Second, if there were no bound on this movement speed, then the system would approach the target exponentially quickly. An exponential approach to target means that irrespective of the starting and end points in acoustic space, the time required to move between phonemes (i.e. the phoneme rate) would be determined in time constants of the system rather than in an absolute time.

To address these two issues, we introduce a limiter, L, that caps the maximum possible absolute change in the control signal by α:

$$L = \min\left(\frac{\alpha}{\max\left(|\boldsymbol{c}'_{fb}|\right)}, 1\right) \tag{4.7}$$

$$\boldsymbol{c}_{fb} = L\boldsymbol{c}'_{fb} \tag{4.8}$$

While L may appear complicated, it is simply checking the rate of change of the fastest moving of all the articulators against α. Then L scales the speeds of all the articulators uniformly so the fastest moving articulator remains within the acceptable limit set by α.

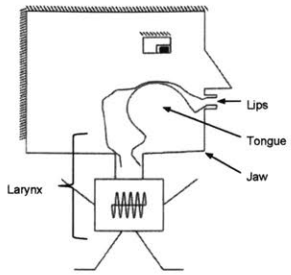


Figure 4.12: The Maeda Man: a schematic version of the articulatory speech synthesis system used as the plant, g, to convert articulator positions to formant frequencies [56, 86].

The Vocal Tract Model

The plant, g, is a function that takes as input articulator positions and outputs formant frequencies. We base our model after a popular model of articulatory speech production, the Maeda model [86] and translated into MATLAB [56]. As implemented, the model allows control of several articulator parameters: jaw height, tongue position, tongue shape, tongue body apex, lip height, lip protrusion, larynx height, and degree of nasal coupling. For our purposes, we assume the nasal coupling parameter has negligible impact, so we have seven controllable articulator parameters. A stylized visualization of the vocal tract model is shown in Figure 4.12.

The full Maeda model is a nonlinear mapping between articulator positions and formant frequencies. The articulator positions are used to create a tube model of the vocal tract and then the tube model is converted into a digital filter from which formant frequencies, formant bandwidths, and formant amplitudes are derived.

In order to allow for fast, iterative updates we approximated the Maeda model as a second order polynomial whose terms are the articulator position terms including their pair-wise cross products. Offline, the true Maeda model was sampled from its parameter space to create a lookup table between articulator positions and formant outputs[3]. The coefficients to the second order polynomial were then fit to the lookup table data using ordinary least squares regression.

■ 4.3.4 Optimization

At this point, we have obtained $\boldsymbol{d}$ (the true phoneme durations) and $\boldsymbol{p}$ (the sequence of phoneme targets including both the phoneme itself and its acoustic representation using the first and second formants). We also have $f(\alpha, w; \boldsymbol{p})$ (our model of the brain), so we can solve for α^* (the go signal strength) and w^* (the auditory width parameter)

[3] genArt2formantMaedaMatrix.m created the full lookup table

in our optimization objective equation (eqn. 4.3) that we repeat here for convenience:

$$\alpha^*, w^* = \text{argmin}_{\alpha,w}\text{Error}\left(f\left(\alpha, w; \boldsymbol{p}\right), \boldsymbol{d}\right). \tag{4.9}$$

With only two parameters over which to optimize, we used a brute force grid search over the parameter space. The limits of the parameter space were set heuristically from initial experiments. We explore 10 values for each parameter for a total of 100 unique pairs. The (α, w) pair that resulted in minimum error was taken as the optimum pair, (α^*, w^*). Our error function is the mean absolute error between measured and true phoneme durations over the first one hundred phonemes from each of the read passages. Among these one hundred phonemes, we exclude from the error computation all phonemes whose true duration was greater than 200 ms because such a long duration implies other forms of timing control may be in operation.

■ 4.4 Results

Figure 4.13 shows the algorithm output for one speech segment[4]. The predicted phoneme duration is on the y- axis and the actual measured phoneme duration is on the x-axis.

Figure 4.14 shows a scatter plot of each of the subjects parameterized by each subject's estimated (α, w) pair for the respective read passages. Compare this figure to Figure 4.5. Qualitatively, these two features do not appear to segregate the subject groups. Depressed subject have a slightly larger w and a smaller α. The effect is more pronounced along the w-axis than the α- axis.

We report the group level statistics in Table 4.4, 4.5, 4.6[5]. These group level statistics quantify the poor separability of the depressed and control classes for each of the read passages on the basis of the neural rate features. The goodness of fit is the mean absolute error in phoneme duration. Overall fit of the model appears reasonable given that the median mean absolute error across the passage for both subject groups is always less than 41 milliseconds. However, these model parameters are fit per utterance in which case a baseline measurement for each utterance would simply be the mean phoneme duration within the utterance as a simple "model" to fit the data.

As a baseline for model fit, we can assign to each phoneme a duration equal to the mean phoneme duration within the utterance for that individual. When we compute the mean absolute error using this baseline metric, the MAE is smaller than when we use the model predicted phoneme durations.

MAE, while a reasonable metric, does not tell the complete story. We also computed the Spearman correlation between the predicted phoneme durations and the actual phoneme durations. The optimization was still performed over minimizing MAE, but we analyzed the Spearman correlation to test whether the dynamic changes of the phonemes was being captured by the w and α parameters. For comparison, if the

[4] abcRate_scatter.ipynb
[5] abcRate_analysis.ipynb

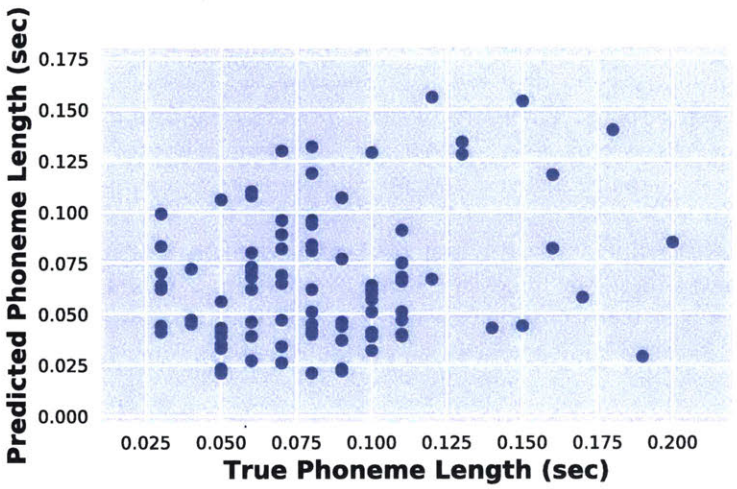


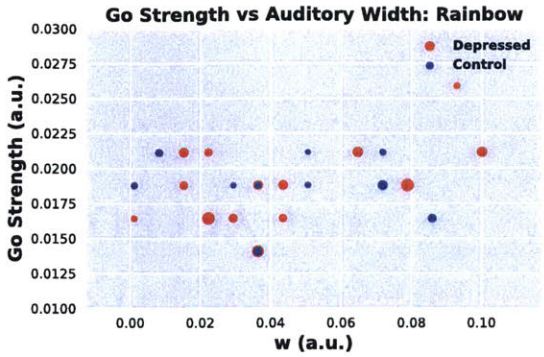
Figure 4.13: Simulated and actual phoneme durations for one emotional sentence for one subject with depression. Spearman correlation: 0.18 ($p = 0.08$). Mean absolute error: 37 ms.

Table 4.4: Phoneme rate feature and goodness-of-fit statistics (MAE): Rainbow.

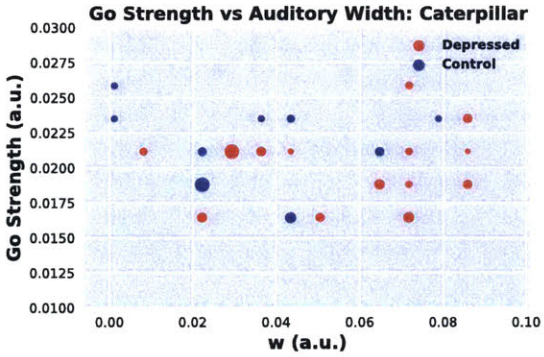
	w		α		Goodness-of-Fit	
	Median	IQR	Median	IQR	Median	IQR
Group						
Control	0.04	0.04	0.019	0.002	0.034	0.005
Depressed	0.03	0.05	0.019	0.003	0.037	0.006

Table 4.5: Phoneme rate feature and goodness-of-fit (MAE) statistics: Caterpillar.

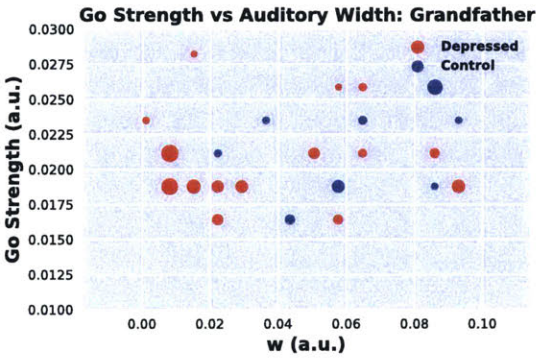
	w		α		Goodness-of-Fit	
	Median	IQR	Median	IQR	Median	IQR
Group						
Control	0.04	0.04	0.022	0.002	0.037	0.005
Depressed	0.05	0.04	0.021	0.002	0.037	0.005



(a) Rainbow



(b) Caterpillar



(c) Grandfather

Figure 4.14: Scatter plot of subjects using the phoneme rate model parameters for each passage. Dot size is proportional to goodness-of-fit (smaller dot implies better fit i.e., smaller mean absolute error). a.u. = arbitrary units.

Table 4.6: Phoneme rate features of auditory width, w, go signal strength, α, and goodness-of-fit (MAE) statistics: Grandfather.

	w		α		Goodness-of-Fit	
	Median	IQR	Median	IQR	Median	IQR
Group						
Control	0.06	0.04	0.022	0.005	0.037	0.006
Depressed	0.05	0.04	0.021	0.002	0.040	0.009

Table 4.7: Phoneme rate model goodness-of-fit using Spearman correlation between actual and predicted phoneme durations.

		Mean	Std. Dev.	Median	IQR
Protocol	Group				
Rainbow	Control	0.23	0.07	0.22	0.07
	Depressed	0.17	0.09	0.16	0.12
Caterpillar	Control	0.13	0.06	0.13	0.06
	Depressed	0.19	0.09	0.17	0.14
Grandfather	Control	0.11	0.09	0.09	0.08
	Depressed	0.13	0.10	0.13	0.20

average phoneme duration were used as the baseline, the Spearman correlation would be zero.

Table 4.7 shows the phoneme duration Spearman correlations in tabular form[6]. Absolute performance numbers are between 0.11 and 0.23 for the mean Spearman correlation coefficient for both groups for all passages. There is also a range of variation in terms of subjects for whom fit was stronger vs weaker, and there was a range of variation across the passages. Overall correlations are highest for the Rainbow passage for both groups of subjects. Correlations for the Grandfather passage and Caterpillar passage are similar but lag the correlations of the Rainbow passage.

■ 4.5 Discussion

This chapter introduced a neurocomputational model of phoneme rate variability. The model leveraged two key parameters within a neurally inspired control framework, the DIVA model. This chapter quantified the ability of these parameters introduced by Guenther [59] to fit a model of phoneme rate variability to read speech.

These features were hypothesized to provide a compact representation of unseen dynamics of speech production and neural influences in order to discriminate between depressed and control subjects on the basis of how these individual reacted. Upon an

[6] abcRate_analysis.ipynb

analysis of the separability of classes, we did not see predictive value in using these features and did not progress to constructing a classifier. At best, we might say depressed subjects show more variability overall in the auditory width parameter rather than the go signal strength parameter when comparing the distributions within the Grandfather and Caterpillar passages.

The underlying basis for discriminating between the two classes was that the parameters themselves would at least model phoneme rate variability well. In this aim, we have had more success. While these correlations are not large, ranging between 0.11 and 0.23, they are encouraging. They show that a compact model of control following the principles of the amplitude strategy and the velocity strategy is able to capture a certain level of variability in how subjects traverse articulatory space and how that traversal relates to phoneme durations.

In reflecting on these results, we return to the idea that speech rate control takes place at multiple levels: at the passage, the sentence, the word, and the phoneme. Our model assumed rate control operating only at the phoneme level as a bottom up way of explaining phoneme rate variability. However, effects of speech rate can occur at multiple levels. For example, two phonemes commonly produced in sequence as a syllable may have an optimized trajectory in acoustic space different than the one found by the model. Furthermore, the algorithm treated both vowels and consonants similarly in that each was a point in acoustic target space that could be obtained. Because our model did capture some measure of variability, we have provided a means of experimental support for treating phonemes as targets in acoustic space. However, subjects may approach production of consonants and vowels differently given that consonants generally have strict timing requirements relative to vowels. Therefore, an auxiliary constraint based on the class of the phoneme (e.g., vowel or consonant) may be controlled. In the limit of this idea, there may be phoneme dependent auditory width targets depending on the density of phonemes in the relevant region of acoustic space as well as phoneme dependent go-strengths.

In our scatter plots of the parameters, we saw greater variance along the auditory target dimension than the go signal strength. We propose two different experiments to investigate this phenomenon. First, a direct measurement of articulatory speed could be made as subjects hit acoustic targets. Such work was performed non-invasively using real time MRI by Lammert et al. [77]. Another possibility is to use electromagnetic articulagraphy (EMA). EMA places small magnets on the articulators which allow tracking their position as a subject speaks. While more invasive than a real time MRI measurement, this modality facilitates point tracking of articulators allowing additional confirmation of the amplitude or velocity strategy.

To investigate auditory target differences and whether subjects are moving these targets or are changing the width of the targets, we propose obtaining an independent measure of auditory acuity for different phonemes. Furthermore, we could use a different neuroimaging technique to identify if subjects have a narrow auditory target and are unwilling to remedy a registered error or in fact are oblivious to the error itself. One

way to investigate this idea would be to record a subject's speech and play it back to them with shifted formants while measuring for a mismatch response. Depressed subjects may register just as large or small a mismatch response as control subjects (i.e. they are just as sensitive to perturbations of their speech), but depressed subjects may lack the motivation to correct such a mismatch. This error response, while not deriveable from overt speech, could be a novel, reliable biomarker that uses a nearly passive protocol. Passive protocols are preferred as they are less likely to be influenced by subjects consciously or unconsciously changing their speech patterns when they know they are being assessed for depression.

Chapter 5

Phoneme Rate and Acoustic Biomarkers of Depression

THUS far we have focused on phoneme rate as a biomarker and neurocomputational features derived from phoneme rate. In this chapter we return to the motivating goal of tracking depression through vocal biomarkers. In addition to these features there are several other model based and model free features that can be extracted from the voice. We briefly describe these other features, and then compare depression prediction performance among the different feature classes. Our aim is two fold: create the best possible depression prediction system, and compare the value of model based features to model free features.

Section 5.1 provides an overview of speech and depression research. Sections 5.2, 5.3, and 5.4 walk through the method, results, and discussion of a voice-based, depression prediction system.

■ 5.1 Speech and Depression

Section 5.1.1 summarizes the state of the art methods applied to using speech to assess depression and their success. Section 5.1.2 describes a framework in which speech features and their associated elicitation protocols can be viewed from a neurological control point of view. In other words, we can view speech production as a complex system whose function under different circumstances gives rise to various attributes. These features can provide insight into how subcomponents of the system are functioning.

■ 5.1.1 Characteristics of Depressed Speech

Speech has been investigated for its utility in diagnosing psychological pathologies for years. While we will focus on the history of speech and depression [34], speech changes have been associated with many other neurological disorders, including Parkinson's disease [131] and traumatic brain injury [107]. In this section, we will review the features generally extracted from speech and current performance associated with the best features. For a thorough review, see Cummins et al. [33].

While there are countless different features that can be extracted from speech, they

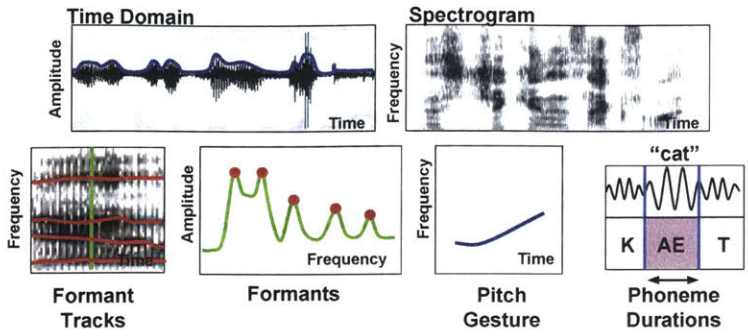


Figure 5.1: Examples of low-level features that can be derived from the speech waveform. The large diversity of measurements makes speech an information rich biomarker.

can be loosely organized into low level features of the articulation process and high level cognitive. The low level feature categories are glottal features, vocal tract features, and prosodic features, and the high level categories are content features and conversational features. Figure 5.1 shows a selection of these features which we will discuss.

Glottal features are characteristics derived from the glottal waveform. The glottis is the opening between the vocal folds through which air passes from the lungs to the vocal tract. The glottal waveform is the volume of air per unit of time (or its derivative) that passes through the glottis. During voiced speech, this will have a periodic structure with each period have a characteristic shape. Unfortunately, the glottal waveform is not directly observable in speech because it is filtered by the vocal tract. Therefore, an inverse filter based on an estimate of the vocal tract transfer function is necessary [91]. Even with a glottal waveform, feature estimates can be noisy because features are generally based on timing and duration of the glottal pulse's waveform, and the vocal folds do not rigidly open and shut but instead undergo a wave-like motion.

Vocal tract features include formant characterization, phoneme rates, measures of articulatory coordination, and forms of energy estimation. The first three formants are typically considered but sometimes the fourth and fifth might be studied too. Formants are characterized by their center frequencies, some measure of dispersion such as the 3 dB bandwidth, and possibly the amplitude if the waveform has been normalized for intensity across subjects or the relative amplitude between formants. phoneme rate measures the average duration of various phonemes and have been discussed at length in this work and elsewhere [129]. Mel frequency cepstral coefficients (MFCC) and its variants, delta MFCC (the first derivative), delta-delta MFCC, and shifted MFCC, transform the speech waveform into a domain that has had exceptional success in all areas of speech signal processing [32]. Articulatory coordination features measure the variability across different speech features such as across formants [143]. Spectral energy features compute the energy in each subband of the speech, and quantify the distribution of energy in the spectrogram [32]. Direct measures of total energy are also used, such

as the Teager Energy Operator [84].

Prosodic features quantify the emotional affect carried by the speech as well as syntactic information and semantic clarification. They have historically focused on the fundamental frequency, and its variability [48]. However, perceptual quality in the form of harmonic to noise ratio, shimmer (amplitude modulation), and jitter (frequency modulation) are also commonly tried.

Conversation features are features that examine the content and flow of free speech, either as a monologue or dialogue. Semantic and lexical features examine the topics of the conversation, and the sentence structure. Speaker to speaker interactions can be measured in terms of turn taking frequency, and pauses between turns. The high level features in this paragraph are outside the scope of the thesis but are mentioned for completeness.

Having reviewed the depth of features that may be extracted from speech, we will summarize how the best algorithms have done at classifying patients as depressed or healthy controls, and the more challenging task of predicting severity of depression. Mundt et al. [93] quantified the difference among responders and non-responders to depression treatment after six weeks. They found significant differences using speaking rate, fundamental frequency coefficient of variation, total recording length, total pause time, and number of pauses. Trevino et al. [129] achieved a class RMSE of 1.24 where each class represented a span of 5 HAMD points using phoneme rate features in a Gaussian Mixture Model. Moore II et al. [91] achieved classification accuracies over 90 percent using a combination of prosodic, vocal tract, and glottal features but did not report a specific feature or features for these categories that was common in their classification tests. [143] achieved an RMSE of 7.42 and a mean absolute error (MAE) of 5.75 when estimating the Beck Depression Inventory II score using articulatory coordination features among the first three formant tracts and among the delta mel cepstral coefficients.

We point out that within this recounting of prior art, the reader does not have an easy time actually comparing performance between algorithms. Some papers perform classification while others perform regression. Some predict HAMD while others predict the BDI. Almost every group uses their own research set which invariably has a unique distribution of depression severity. Therefore, we recommend always at least reporting F_1 score if classification and fraction of variance explained if performing regression.

We have reviewed a subset of these features, many of which are tried if not actually used in each research group's final system. The speech and depression field, while several decades old, still has not converged on a set of best features. Researchers will typically try as many features as can be easily coded and let machine learning algorithms sort out which features were actually useful. While this *ad hoc* procedure sometimes yields acceptable performance, what works and what does not still appears quite variable. This procedure throws into relief the need for an understanding of speech and depression in order to guide feature selection and understand why certain features are successful. We believe model based features help meet this need by providing a

principled, interpretable mechanism for feature generation. In the next section, we interpret some of these features in the context of the speech production process.

■ 5.1.2 Features from a Neurocomputational Modeling Perspective

We present an engineering view of speech control and map different speech tasks onto this view. While speculative, this view provides an organizational framework with which to understand speech features. Broadly, speaking involves one cognitive component and one motor component. The cognitive component can be concept planning, sentence structure (which includes linguistic structure as well as prosodic modulation). The motor component includes a feedforward component and feedback component.

Sentence production, picture naming, and non-word repetition draw upon different levels of a hierarchical planning process. At the lowest level, heard non-word repetition is a reproduction of an auditory target. Read non-words requires synthesis of legal but presumably infrequently used phoneme combinations. Picture naming requires chunk retrieval but not synthesis. Free speech sentences (e.g. question answering) require planning content for a response, and a full production encompassing syntactic constraints and prosodic modulation (both linguistic and potentially affective prosody). Read sentences require imputing the specific linguistic and explicit or implicit affective prosody. Analysis of phoneme rates from sentences and non-words can determine context effects on production and sub-unit synthesis effects. Response times for picture naming signal retrieval and conversion of concept to speech motor program.

Tongue twisters and diadochokinetic tasks such as fast repetition of "pa-ta-ka" stress feedforward motor execution. From these tasks, we may extract features of formant trajectories as the objective is to rapidly change between articulatory targets, and rapid changing of the vocal tract yields rapid changes in formants. These also lend themselves to studying the locations of vowels (the quasi steady state formants for vowel phonemes), the quality of consonant production, the duration of vowels and consonants, and the change between voicing and unvoiced segments of speech.

Sustained phonations are well suited for feedback execution as the target is not time varying but nonetheless needs to be maintained in the presence of motor execution noise or external experimental perturbations (perturbations that may not rise to the level of conscious awareness). From sustained phonations we can derive measures of voice quality and voice control because the nominal objective is to maintain as constant an acoustic production as possible. Voice quality, which is characterized by breathiness, strength, and glottal flow features, provide additional insight into the biomechanical properties of the speech source system that includes the respiratory system and the vocal folds.

■ 5.2 Materials and Methods

The model based features are the neural computational phoneme rate features introduced in the previous chapter as well as neurocomputational source (NCS) and

Table 5.1: Speech features from a neurocomputational point of view.

Speech Phase	Protocol	Speech Motor Control Characteristic	Acoustic Features
Planning	Sentences (read, and free), picture naming	Content and motor program retrieval, long term planning within syntactic and prosodic constraints	Speech rate, f0 contours, intensity contours, context dependent phoneme rate
Planning	Nonword repetition (heard and read)	Sequence effects	phoneme level rates, time to onset of production
Feedforward Motor	Tongue twister, diadochokinetic ("pa-ta-ka")	Fine motor control properties	Articulatory coordination via formants. Speech rate, phoneme rate, articulation rate, vowel space area, voice onset/offset times, amplitude and dynamics of formant trajectories, vowel space area
Feedback Motor	Sustained vowel phonation (with and without perturbations)	Biomechanical vocal source properties	Voice quality: HNR, CPP, amplitude, glottal flow features
Feedback Motor	Speech Perturbations	Feedback control properties	Formant and f0 stationarity. Response times, directions, and magnitudes of compensations to perturbations

neurocomputational tract features (NCT). The NCT features were first introduced by Williamson et al. [145]. These features are the coordination features derived from the positions of articulators inferred through the DIVA model on the input formant tracks. By analogy, we created a neurocomputational model of vocal source control in Ciccarelli et al. [29]. The neurocomputational source features are the coordination features derived from inferred vocal source muscle activations, specifically the cricothyroid and thyroarytenoid. The control framework for the NCS model follows the DIVA control architecture, but we replaced a plant that converted articulator positions to formants with a plant that converted vocal source muscle activations to a fundamental frequency.

We now provide an extended example of a vocal source model developed for a held [AA] vowel in depression. This work was first described in Ciccarelli et al. [29].

■ 5.2.1 Vocal Source Model

A novel contribution of this research is a feature set derived from a biophysically inspired model of the vocal source and computationally plausible neural control mechanism. We sought a model of the vocal folds and their control mechanism in order to derive unobserved but existent muscle activations in the larynx. This approach is similar to, but substantially more developed than, Williamson et al. [145] in that we use the same control system paradigm, but differs in that we focus on the vocal source rather than vocal tract, and we introduce a biophysically inspired model of the vocal source. We defer a discussion of the broader implications of this approach until later, and here focus on implementation.

Control Framework

We adopted the neurocomputational control scheme hypothesized by Guenther et al. [61] in the Directions into Velocities of Articulators model. The model, adapted for vocal source control, is shown in Figure 5.2. In this scheme, there is an auditory target, which we set as the extracted fundamental frequency time series of the input speech. The forward model transforms the unobserved motor space parameters to the observed auditory space. A second component, termed the inverse model, transforms the error between the observed auditory component and the produced auditory signal into a feedback motor update. In the system used for this work, we have omitted the auditory estimation step and the somatosensory feedback as unnecessary for illustrating the central theme of using a neurocomputational model to extract features.

Implementation

The system is trained in an iterative process to determine the latent parameters necessary to reproduce the auditory target. When there is a sufficient match between the auditory target and the model production, the latent parameters are assumed to be representative of the true latent parameters of the system. In our implementation, we have a one dimensional auditory target, the fundamental frequency, and we have

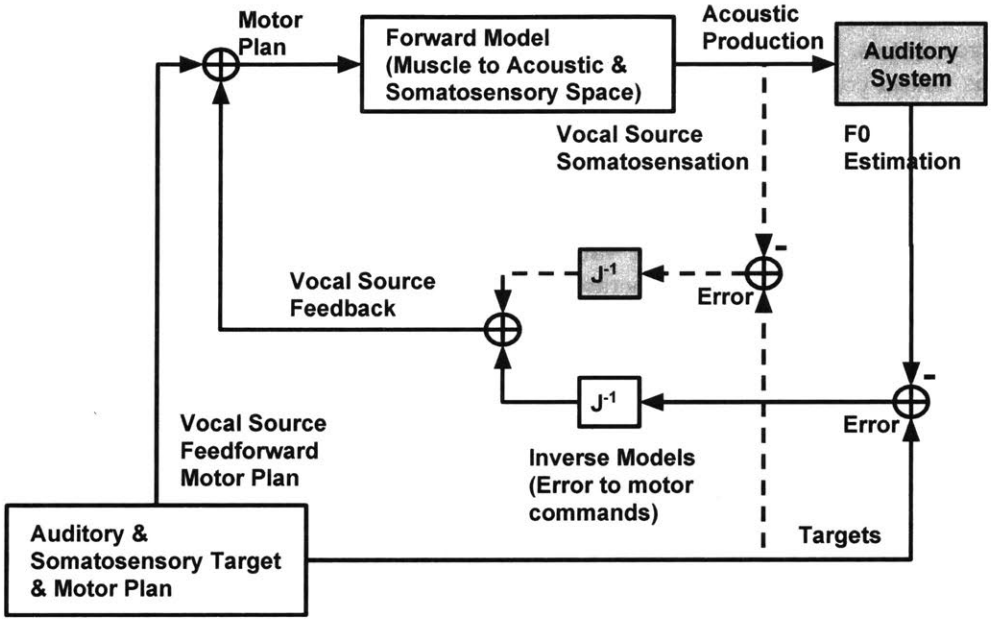


Figure 5.2: Neurocomputational control framework for the vocal source. The biophysical source model enters in the forward model and auditory inversion blocks. Dotted lines and gray modules are not used in the results.

a two dimensional latent space that nominally represents the neural activation to the cricothyroid (CT) and thyroarytenoid (TA) muscles of the larynx.

The CT and TA muscles along with subglottal pressure and other intrinsic laryngeal muscles influence fundamental frequency [126], but we focus on the CT and TA muscles to capture their dominant influence while maintaining tractability. The CT and TA muscles are innervated by the superior laryngeal nerve and recurrent laryngeal nerve respectively, and both nerves are branches of the tenth cranial nerve whose nucleus is in the brainstem [148]. In the context of the perception-action model, we understand the neural signals to the CT and TA muscles as the net contribution of a planned fundamental frequency trajectory determined by prefrontal cortex, limbic system, and basal ganglia integration, and corrective actions based on auditory and somatosensory error signals. The muscles, acting as the motor system and our forward model, translate the neural commands by their interaction with the airstream into a new glottal flow waveform. The new muscle state and the acoustic consequences are perceived and compared to the plan, and the neural activations are updated as needed.

Our forward model is inspired by the Titze and Story biophysical model [125], and we use its computational implementation and extension by Zañartu [149]. We created a mapping of the CT and TA values to the fundamental frequency estimated by a peak picking algorithm of a generated glottal flow waveform for a generic laryngeal system. We approximated the mapping with a quadratic polynomial to quickly evaluate produced fundamental frequency for given CT and TA values because solving the differential equation governing glottal flow is computationally infeasible for seconds worth of speech. Furthermore, with a closed form, differentiable forward model, we were able to quickly compute an inverse of the forward model. We took partial derivatives of the fundamental frequency with respect to the two muscle activations to create a Jacobian matrix, and used the Moore-Penrose inverse of a matrix ($\mathbf{A}^{\dagger} \equiv (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$) in MATLAB (Natick, MA) to create a pseudo-inverse. The pseudo-inverse of the Jacobian converts error signals in sensory space to corresponding motor changes.

We used Praat to extract the fundamental frequency (f_0) trajectory from the vowel waveform Boersma et al. [14], shifted it to our model's f_0 range, and input the trajectory to our vocal source model to infer the hidden CT and TA activation time series.

Muscle Activation

An example of the fundamental frequency extracted from the waveform (True), the model-generated fundamental frequency (Model), and the inferred CT and TA muscle activations are given in Figure 5.3. We acknowledge the strong correlation between the inferred CT value and the fundamental frequency and argue that this is reasonable. Mathematically, the correlation occurs because the gradient of the fundamental frequency surface is strongly aligned with CT, so small changes in muscle activation will impact fundamental frequency most when the CT muscle is changed. The gradient's alignment is consistent with computational simulations of the two muscles in [126] and is consistent with vocal source physiology.

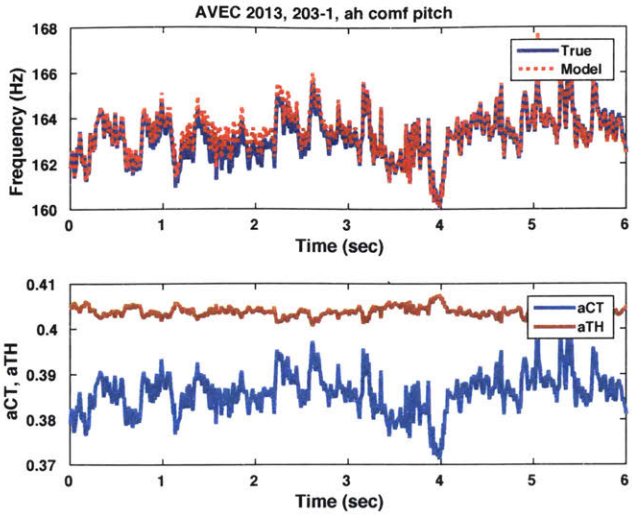


Figure 5.3: True and model-generated fundamental frequency (top) and inferred CT and TA muscle activations (bottom).

The CT, when tensed, pulls the cricoid and thyroid cartilage to directly tense the vocal cords and therefore increase fundamental frequency. The CT has a larger influence than the TA on changes in fundamental frequency because the CT directly regulates vocal fold tension whereas the TA has an indirect effect. To understand the difference in magnitude of influence of the two muscles, we appeal to the body cover model of the vocal folds that describes the vocal folds as a muscular body loosely connected to a covering with different mechanical properties than the body. The TA may differentially slacken the cover component of the model while increasing the tension of the body. Depending on the net tension increase or decrease of the body cover system, the fundamental frequency may increase or decrease [78, 127]. *Ex vivo* laryngeal stimulation experiments confirm the dominant role of the CT muscle and the nuanced role of the TA muscle [27].

Feature Extraction from Muscle Activations

From the CT and TA time series, we generate a multi-scale set of features based on the cross correlation of these waveforms. The features that are used in the depression prediction system are the eigenvalues from a set of matrices whose elements are samples from the auto and cross correlations of CT and TA. This technique was introduced by Williamson et al. [142], and we summarize the main points of the procedure here.

Let $\mathbf{Z}$ be a n_s by n_c matrix where n_s is the number of samples in each input waveform, and n_c is the number of input waveforms. Construct a new matrix $\mathbf{X}$ as the horizontal concatenation of time delayed versions of $\mathbf{Z}$:

$$\mathbf{X} = \left[\mathbf{Z}_{d_0}, ..., \mathbf{Z}_{d_k}, ..., \mathbf{Z}_{d_{K-1}}\right] \tag{5.1}$$

where $\mathbf{Z}_{d_k}$ is a d_k sample delayed version of $\mathbf{Z}$. Practically, this means pre-pending d_k rows of zeros to $\mathbf{Z}$ and removing the last d_k rows of $\mathbf{Z}$ such that the delayed matrix is still of dimension n_s by n_c.

The individual delays, d_k, are constructed as:

$$d_k = k \cdot \delta_j \text{ for } k = 0, ..., (K-1) \tag{5.2}$$

where δ_j is the step size for the j^{th} delay scale, and K is the total number of delays. If the waveform is delayed by more than the number of samples in the waveform, then the delayed matrix will be all zeros, so the following inequality should be kept

$$(K-1) \cdot \max_j \delta_j \leq n_s. \tag{5.3}$$

In practice, the number of delays are chosen as the square root of the number of samples, and then the delay scales, δ_j are chosen as powers of 2 until equation 5.3 is violated.

The next step is to create a biased covariance matrix, $\mathbf{C}$, and a biased correlation coefficient matrix, $\mathbf{R}$, using $\mathbf{X}$:

$$\mathbf{C} = \frac{1}{n_s} f(\mathbf{X})^T f(\mathbf{X}) \tag{5.4}$$

$$\mathbf{R} = \frac{1}{n_s} g(\mathbf{X}) g(\mathbf{X}) \tag{5.5}$$

where f removes the column-wise mean from each column of $\mathbf{X}$ and g removes the column-wise mean from each column of $\mathbf{X}$ and then normalizes each column to unit variance.

Finally, the eigenvalues of these two matrices are rank sorted and used as feature vectors. An additional two features, the log of the sum of the eigenvalues $\mathbf{C}$ and the log of the product of the eigenvalues of $\mathbf{C}$ can also be extracted.

The cross-correlation features are applied to the articulatory and acoustic trajectories derived from the phoneme rate model in Chapter 4, but we also use the phoneme rate model α, w, and goodness-of-fit statistics as their own feature set.

■ 5.2.2 Model Free Features

All Phoneme Statistics

In Chapter 2 we defined phoneme rate as the reciprocal of the mean phoneme duration, and we mentioned that mean phoneme duration is one summary statistic for the distribution of durations for each phoneme. To the collection of mean phoneme durations, we add a variety of other distribution measures, which we collectively call "All phoneme Statistics". These other phoneme stats include the variance, number of times

the phoneme occurrence ("count"), the rate (reciprocal of the mean duration), the interquartile range of the duration, the minimum and maximum duration, the duration range (maximum duration minus minimum duration), standard error of the mean, 25^{th} and 75^{th} percentiles, and percentiles (10^{th} through 90^{th} percentiles in increments of ten percent).

Stevens [124] introduced a class of features that are a generalization of the phoneme. These features, like phonemes, discretize the speech waveform. However, they are based on acoustic landmarks meant to emphasize acoustic contrast between segments. Each generalized phoneme is described by a binary feature vector that includes descriptors such as vowel, consonant, continuant (oral vocal tract is completely closed, e.g. [F]), or sonorant (non-turbulent airflow e.g. [M]).

In the spirit of using these binary features as an alternate representation of discrete segments, we define several new classes of "phonemes"[1]. A phoneme class is a group of phonemes that share a similar manner of articulation. The phoneme classes are: vowels, consonants, voiced, unvoiced, front, central, and, back vowels, high, middle, and low vowels, and stop, nasal, fricative, approximant, and affricate consonants. Table 5.2 provides a complete listing of memberships of each of the phonemes to these different classes[6, 83]. Following Lee and Hon [80], we group [AX] and [AH] together as [AH], but following Livescu [83] we ascribe to [AH] the properties of [AX]. We reasoned that phoneme classes may provide greater data support for particular motor actions that would be sensitive to depression, and therefore provide a metric with greater signal to noise ratio then when the class's constituent phonemes are considered on their own. This is similar to Trevino et al. [129] who aggregated phonemes by some of these classes to look for patterns among common discriminative phonemes.

Finally, we include the speaking rate, the articulation rate, the total duration of the utterance including silence, the total duration excluding silence, the total number of phonemes, and the total number of unique phonemes within the utterance. All told, each utterance has 1216 phoneme statistics including the mean phoneme duration which was used in Chapter 2.

Opensmile

Opensmile is an omnibus, speech feature extraction software [41]. It implements many of the features discussed earlier in Section 5.1. We use it here as the quintessential package for covering a large breadth of speech features in a manner that is highly reproducible for other researchers as the software is readily available for academic research. Opensmile uses configuration files to generate sets of features, some of which were used for various speech signal processing challenges. We use the emo_large.conf and ComParE_2016.conf.

Opensmile features are generated similar to how we discussed our phoneme statistics. A set of low level features are enumerated, and then summary statistics on the low level feature descriptors are applied. There are 39 summary statistics for each low level feature including the feature distribution's centroid, range, variance, standard

[1] vw_co_classes_to_latex.ipynb

Table 5.2: Assignment of phonemes to different phoneme classes. 1 means membership of the class denoted by the column. After [6,83]

Phoneme	Voiced	Voiceless	Vowel	Consonant	High	Middle	Low	Front	Central	Back	Stop	Nasal	Fricative	Approximant	Affricate
IH	1	0	1	0	1	0	0	1	0	0	0	0	0	0	0
EH	1	0	1	0	0	1	0	1	0	0	0	0	0	0	0
AE	1	0	1	0	0	0	1	1	0	0	0	0	0	0	0
AA	1	0	1	0	0	0	1	0	0	1	0	0	0	0	0
AH	1	0	1	0	0	1	0	0	1	0	0	0	0	0	0
UH	1	0	1	0	1	0	0	0	0	1	0	0	0	0	0
AO	1	0	1	0	0	1	0	0	0	1	0	0	0	0	0
ER	1	0	1	0	1	0	0	0	1	0	0	0	0	0	0
IY	1	0	1	0	1	0	0	1	0	0	0	0	0	0	0
EY	1	0	1	0	0	1	0	1	0	0	0	0	0	0	0
OW	1	0	1	0	0	1	0	0	0	1	0	0	0	0	0
UW	1	0	1	0	1	0	0	0	0	1	0	0	0	0	0
AY	1	0	1	0	0	0	1	0	1	0	0	0	0	0	0
OY	1	0	1	0	0	1	0	0	0	1	0	0	0	0	0
AW	1	0	1	0	0	0	1	0	1	0	0	0	0	0	0
P	0	1	0	1	0	0	0	0	0	0	1	0	0	0	0
B	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0
T	0	1	0	1	0	0	0	0	0	0	1	0	0	0	0
D	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0
K	0	1	0	1	0	0	0	0	0	0	1	0	0	0	0
G	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0
CH	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1
JH	1	0	0	1	0	0	0	0	0	0	0	0	0	0	1
F	0	1	0	1	0	0	0	0	0	0	0	0	1	0	0
V	1	0	0	1	0	0	0	0	0	0	0	0	1	0	0
TH	0	1	0	1	0	0	0	0	0	0	0	0	1	0	0
DH	1	0	0	1	0	0	0	0	0	0	0	0	1	0	0
S	0	1	0	1	0	0	0	0	0	0	0	0	1	0	0
SH	0	1	0	1	0	0	0	0	0	0	0	0	1	0	0
Z	1	0	0	1	0	0	0	0	0	0	0	0	1	0	0
ZH	1	0	0	1	0	0	0	0	0	0	0	0	1	0	0
HH	0	1	0	1	0	0	0	0	0	0	0	0	1	0	0
R	1	0	0	1	0	0	0	0	0	0	0	0	0	1	0
M	1	0	0	1	0	0	0	0	0	0	0	1	0	0	0
N	1	0	0	1	0	0	0	0	0	0	0	1	0	0	0
NG	1	0	0	1	0	0	0	0	0	0	0	1	0	0	0
L	1	0	0	1	0	0	0	0	0	0	0	0	0	1	0
Y	1	0	0	1	0	0	0	0	0	0	0	0	0	1	0
W	1	0	0	1	0	0	0	0	0	0	0	0	0	1	0

deviation, kurtosis, 25, 50, and 75^{th} quartiles, the differences between the quartiles, and the 95^{th} and 98^{th} percentiles. Low level features include thirteen mel-frequency cepstral coefficients (MFCCs) and also delta MFCCs and delta-delta MFCCs. There are also measures of fundamental frequency, the fundamental frequency envelope, the zero crossing rate, and the distribution of energies within the frequency spectrum (e.g., measures of spectral roll off or how fast energy per frequency band declines at higher frequencies).

To partially disambiguate between the importance of many summary statistics vs the importance of the underlying features, we report results using just the "amean" summary statistic which is the Opensmile descriptor for arithmetic mean as well as all the summary statistics ("all"). We also apply the cross-correlation technique to the acoustic (formant and fundamental frequency) trajectories extracted directly from each acoustic waveform. A description for interpreting the feature set names is in Table 5.3.

5.2.3 Machine Learning

We perform shuffle-split cross-validation folds on each feature set according to the pipeline described in Chapter 2.2.4. Our reported metric is receiving operating characteristic area under the curve (ROC AUC).

5.3 Results

The results from our cross-validation tests are in Tables 5.4, 5.5 and 5.6. Across feature sets, classification performance is best on the Grandfather passage with the best performing feature set being Opensmile mean features with a mean ROC AUC 0.82. The Caterpillar passage and Rainbow passage have peak performances of 0.79 and 0.81, also using some form of the Opensmile feature sets. The additional statistics based on phoneme durations provides a small performance boost over just using the mean phoneme durations. The model based features including the phoneme rate model features do not perform strongly with the exception of the ausRate features for the Caterpillar passage.

5.4 Discussion

We see that the Opensmile features provide the best stand alone classification performance. At first this might suggest that modeling is not needed, and only data. However, "modeling" is more than a particular set of equations. Modeling is an approach in which one incorporates prior knowledge about the underlying process into one's system. This thesis has approached modeling from the beginning as a means of identifying biomarkers of depression. While it has succeeded to a lesser degree with computational models of phoneme rate, a close look at the Opensmile features suggests that they are not as model-free as our initial description implied.

The spectral features that constitute many of the Opensmile features were inspired

Table 5.3: Description of feature sets used for classification.

Feature Set	Description
model_merge	Merge of all model based features: ausRate, gridRate, smTrt, smSrc (CT and TA together)
vote_fc-task-taskname_merge_fcs-deal_all	Merge of all feature sets.
vote_fc-phnhydec_unnorm_fcs-deal_all	Mean durations and higher order statistics of all 40 phonemes and phoneme classes
vote_fc-phnhydec_unnorm_fcs-deal_Mean	Mean durations of all 40 phonemes
vote_fc-voice_anorm_opensmile16_task_fcs-deal_all	Opensmile ComParE_2016.conf, all features
vote_fc-voice_anorm_opensmile_fcs-deal_all	Opensmile emo_large.conf, all features
vote_fc-voice_anorm_opensmile_fcs-deal_aMean	Opensmile emo_large.conf, feature means
vote_fc-voice_ausRate_xcorr_fcs-deal_all	Cross-correlation features derived from auditory and articulator trajectories based on simulating phoneme durations using a DIVA inspired model of phoneme rate control.
vote_fc-voice_gridRate_fcs-deal_all	Phone rate model parameters including α and w as well as goodness-of-fit measures
vote_fc-voice_karma_smTrt_xcorr_fcs-deal_all	Cross-correlation features derived from articulator trajectories generated from DIVA model control of the vocal tract. Based on formants extracted using KARMA.
vote_fc-voice_karma_xcorr_fcs-deal_all	Cross-correlation of the formants extracted using KARMA.
vote_fc-voice_praatPitch_smSrc_xcorrllCT_fcs-deal_all	Cross-correlation features derived from the cricothyroid muscle activation generated from DIVA model control of the vocal source. Based on the fundamental frequency trajectory extracted using Praat.
vote_fc-voice_praatPitch_smSrc_xcorrllTA_fcs-deal_all	Cross-correlation features derived from the thyroarytenoid muscle activation generated from DIVA model control of the vocal source. Based on the fundamental frequency trajectory extracted using Praat.
vote_fc-voice_praatPitch_smSrc_xcorr_fcs-deal_all	Cross-correlation features derived from the cricothyroid and thyroarytenoid muscle activation generated from DIVA model control of the vocal source. Based on the fundamental frequency trajectory extracted using Praat.
vote_fc-voice_praatPitch_xcorr_fcs-deal_all	Cross-correlation features derived from the fundamental frequency trajectory extracted using Praat.

Table 5.4: Classification performance by feature type for the Rainbow passage: area under the ROC curve.

Feature Set	Mean	Std. Dev.	Median	IQR
model_merge	0.46	0.03	0.46	0.04
vote_fc-20170518161533_task-rainbow_merge_fcs-deal_all	0.79	0.03	0.78	0.03
vote_fc-phnhydec_unnorm_fcs-deal_all	0.70	0.02	0.70	0.03
vote_fc-phnhydec_unnorm_fcs-deal_Mean	0.72	0.02	0.73	0.02
vote_fc-voice_anorm_opensmile16_task_fcs-deal_all	0.81	0.02	0.81	0.03
vote_fc-voice_anorm_opensmile_fcs-deal_all	0.79	0.03	0.79	0.03
vote_fc-voice_anorm_opensmile_fcs-deal_aMean	0.76	0.04	0.76	0.07
vote_fc-voice_ausRate_xcorr_fcs-deal_all	0.44	0.06	0.44	0.07
vote_fc-voice_gridRate_fcs-deal_all	0.55	0.04	0.55	0.07
vote_fc-voice_karma_smTrt_xcorr_fcs-deal_all	0.59	0.03	0.59	0.03
vote_fc-voice_karma_xcorr_fcs-deal_all	0.63	0.05	0.61	0.05
vote_fc-voice_praatPitch_smSrc_xcorrCT_fcs-deal_all	0.35	0.03	0.34	0.03
vote_fc-voice_praatPitch_smSrc_xcorrTA_fcs-deal_all	0.38	0.06	0.38	0.07
vote_fc-voice_praatPitch_smSrc_xcorr_fcs-deal_all	0.36	0.04	0.35	0.05
vote_fc-voice_praatPitch_xcorr_fcs-deal_all	0.45	0.05	0.45	0.08

Table 5.5: Classification performance by feature type for the Caterpillar passage: area under the ROC curve.

f_set	Mean	Std. Dev.	Median	IQR
model_merge	0.57	0.03	0.58	0.04
vote_fc-20170518161533_task-caterpillar_merge_fcs-deal_all	0.77	0.03	0.77	0.04
vote_fc-phnhydec_unnorm_fcs-deal_all	0.61	0.05	0.62	0.05
vote_fc-phnhydec_unnorm_fcs-deal_Mean	0.54	0.05	0.54	0.06
vote_fc-voice_anorm_opensmile16_task_fcs-deal_all	0.79	0.03	0.79	0.04
vote_fc-voice_anorm_opensmile_fcs-deal_all	0.77	0.03	0.78	0.01
vote_fc-voice_anorm_opensmile_fcs-deal_aMean	0.74	0.03	0.73	0.03
vote_fc-voice_ausRate_xcorr_fcs-deal_all	0.73	0.04	0.73	0.05
vote_fc-voice_gridRate_fcs-deal_all	0.51	0.04	0.50	0.05
vote_fc-voice_karma_smTrt_xcorr_fcs-deal_all	0.42	0.07	0.42	0.09
vote_fc-voice_karma_xcorr_fcs-deal_all	0.68	0.06	0.68	0.07
vote_fc-voice_praatPitch_smSrc_xcorrCT_fcs-deal_all	0.32	0.07	0.32	0.13
vote_fc-voice_praatPitch_smSrc_xcorrTA_fcs-deal_all	0.19	0.05	0.20	0.04
vote_fc-voice_praatPitch_smSrc_xcorr_fcs-deal_all	0.45	0.05	0.44	0.08
vote_fc-voice_praatPitch_xcorr_fcs-deal_all	0.32	0.04	0.33	0.05

Table 5.6: Classification performance by feature type for the Grandfather passage: area under the ROC curve.

Feature Set	Mean	Std. Dev.	Median	IQR
model_merge	0.33	0.07	0.32	0.10
vote_fc-20170518161533_task-grandfather_merge_fcs-deal_all	0.80	0.04	0.81	0.04
vote_fc-phnhydec_unnorm_fcs-deal_all	0.49	0.03	0.49	0.03
vote_fc-phnhydec_unnorm_fcs-deal_Mean	0.47	0.05	0.48	0.07
vote_fc-voice_anorm_opensmile16_task_fcs-deal_all	0.72	0.04	0.72	0.05
vote_fc-voice_anorm_opensmile_fcs-deal_all	0.81	0.04	0.82	0.04
vote_fc-voice_anorm_opensmile_fcs-deal_aMean	0.82	0.04	0.83	0.05
vote_fc-voice_ausRate_xcorr_fcs-deal_all	0.48	0.06	0.49	0.04
vote_fc-voice_gridRate_fcs-deal_all	0.40	0.05	0.40	0.03
vote_fc-voice_karma_smTrt_xcorr_fcs-deal_all	0.32	0.05	0.32	0.04
vote_fc-voice_karma_xcorr_fcs-deal_all	0.72	0.05	0.72	0.05
vote_fc-voice_praatPitch_smSrc_xcorrCT_fcs-deal_all	0.60	0.05	0.61	0.07
vote_fc-voice_praatPitch_smSrc_xcorrTA_fcs-deal_all	0.41	0.05	0.42	0.08
vote_fc-voice_praatPitch_smSrc_xcorr_fcs-deal_all	0.49	0.06	0.47	0.06
vote_fc-voice_praatPitch_xcorr_fcs-deal_all	0.50	0.06	0.51	0.09

by a neurobiological understanding of the auditory system which has been shown to perform not just spectral processing of sound, but also spectral processing using features derived from unequal parts of the acoustic frequency spectrum. Consequently, though those features may not here have come directly from a computational model, they are nonetheless part of the modeling framework which this thesis has endorsed as useful for performing assessment of neurobiological disorders.

Chapter 6

Conclusion

THIS thesis advances the scientific and practical understanding of voice as a pragmatic sensor and phoneme rate as a biomarker of depression. Here, we summarize the specific details of the thesis and discuss its contributions within a broader context of computational psychiatry and sensor-based monitoring and treatment.

Section 6.1 reviews the specific aims and findings, and Section 6.2 discusses limitations. Section 6.3 considers investigations enabled by this thesis, and Section 6.4 outlines a broad vision for the role of computational models in tracking and treating psychological disorders.

■ 6.1 Review of Specific Aims and Findings

The driving question behind this thesis was, "How well can phoneme rate and an understanding of it be a biomarker for predicting depression severity?" As depression is both widespread and prevalent, an objective, scalable biomarker for tracking the disorder has the potential to positively impact the treatment and lives of many affected individuals. To answer this question, this thesis pursued several specific aims: characterize phoneme rate in read speech (Chapter 2), identify the neural correlate of phoneme rate (Chapter 3), and model individual phoneme rate variability (Chapter 4). Because phoneme rate is just one voice based biomarker, we took the additional step of comparing the performance of several model-based and model-free approaches to predicting depression (Chapter 5).

We characterized the phoneme rate biomarker by analyzing the phoneme rates of subjects with and without depression using three standard read passages. Our results showed that a longer passage with more than a minute of speech provided improved classification accuracy over shorter passages. Consequently, in a device meant to use voice for tracking depression, researchers should aim for a minimum of one minute of continuous speech.

We conducted a task fMRI study in which control and depressed subjects read implicitly emotional sentences aloud to investigate the neural correlate of the phoneme rate biomarker. Our finding of an interaction between depression and speech production, though not articulation rate in general, in the putamen provides neural evidence for using speech as a biomarker of depression.

We modeled phoneme rate variability by proposing an algorithm for estimating two parameters, the auditory width and the go signal's strength, within the DIVA neurocomputational model. We estimated these parameters, and then evaluated the utility of the model parameters as derived biomarkers for distinguishing between depressed and control subjects.

Finally, we evaluated a suite of model based and model free acoustic biomarkers of depression. Our comparison showed on read speech that model based features did as well as model free features but neither added additional classification performance to the other feature set when both were combined. Note: as we observed in Chapter 5, "model free" is a misnomer in the sense that many of the Opensmile features are actually still inspired by models of the speech production process. Thus, even in "model free", data-driven scenarios, models still prove their worth by structuring how one processes raw data streams.

This thesis presented an innovative development and use of neurocomputational models of speech production for computational psychiatry. We focused on phoneme rate in this exposition, but additional supporting work including speech source and speech tract modeling have also been developed to capture additional components of the brain's dynamic control of speech production.

■ 6.2 Study Challenges

We begin by considering some fundamental research limitations to understanding phoneme rate control and depression. These limitations stem from the hierarchical control of phoneme rate.

When we investigated read speech in the scanner vs outside the scanner, we did not see any articulation rate differences. This could mean rate is being governed at different, higher time scales. For example, the read passages were narratives, so a narrative intent that spanned the full read passage may have influenced the articulation rate (i.e., there existed more freedom for the reader to modulate rate in order to emphasize or de-emphasize parts of the story.). Alternatively or in addition, the single sentence syntax in the scanner combined with the metronomic pacing of the four second repetition time may have imposed a rigid, uniform rate at the sentence-by-sentence level inside the scanner.

Another potential confound is the anatomical change in position of the articulators relative to gravity (subjects were speaking while lying down on their back in the scanner as opposed to upright in front of a computer). Collecting speech while lying down without scanning is a needed control experiment to check for this confound. Additionally, the scanner noise itself, though intermittent and not concurrent with speech production, may influence auditory feedback mechanisms and consequently speech production. This problem would only be worse in continuous scanning. Therefore, one may consider some form of active noise cancellation system to substantially reduce scanner noise while maintaining naturalistic auditory feedback.

Given the possibility of this effect, one might ask whether free speech could be collected in the scanner. Unfortunately, continuous acquisition of free speech responses has two fundamental challenges. First, speaking while imaging will cause artifacts due to physical motion of the brain and changes in magnetic susceptibility. Therefore, those volumes will have to be discarded. Second, if a person extemporaneously spoke or read a passage and the brain was imaged only when the person stopped speaking, the only actual brain activity that would be recorded would be that associated with the last six seconds of speaking. Consequently, most of the time in the scanner would either be artifact contaminated brain volumes or lost acquisition time. Another possibility is to image the brain after pauses of words or phrases, but these would need to last for at least one full second for a complete brain acquisition. Unfortunately, one second may be a considerable amount of pause time in what should be fluent speech.

We focused on articulation rate inside the scanner rather than individual phoneme rates, which we explored in Chapter 2, because despite how different phonemes may be differently affected by depression, fMRI does not have the temporal resolution to observe neural control at the millisecond level. Therefore, it is tempting to turn to another modality such as magnetoencephalography (MEG) as a compliment.

While MEG has a disadvantage relative to fMRI in terms of subcortical sensitivity, if the phoneme rate biomarker has a strong cortical component, then MEG could be a suitable imaging modality. MEG offers the advantage over MRI of millisecond temporal resolution. On the other hand, MEG's instantaneous measurements of neural activity introduce a practical challenge not present with fMRI. Specifically, activating facial muscles to produce speech generates substantial electromagnetic activity that introduces severe artifacts into the desired neural signals that are concurrently being generated.

Additionally, MEG is not a silver bullet for understanding speech rate control. As we have emphasized in previous chapters, speech rate control happens at multiple timescales, and speech rate control operates dynamically. It is implausible to think that a subject plans an entire passage or even sentence completely prior to actual execution of the sentence. The online planning nature of speech production makes studying rate control particularly challenging. In Parkinson's disease, noted features of the pathology are a greater acceleration of articulation rate relative to controls and fewer pauses over the course of a read passage[120]. These effects may imply higher order effects than might simply be explained by generalized articulatory movement difficulty which is present in Parkinson's disease.

Another fundamental limitation is on how we evaluate depression. We judged our prediction performance based on the similarity of the automatic prediction result with the self-reported BDI depression score. Unfortunately, that means our algorithms can not be better than this truth metric which could be variable within and between subjects. Ultimately, we would like to perform a prospective trial in which speech could be used to predict which patients would improve from a given treatment strategy, though that form of study also still relies on a metric of depression severity.

A particular challenge of any measure of a complex system such as speech and depression is the heterogeneity among individuals. While we controlled for heterogeneity in speech to a first order by considering English only and matching subjects for age and sex, individual variability still exists. Heterogeneity of a psychological disorder as complex as depression is a different challenge entirely. While this study's power could benefit from more subjects, like any study, the specific reason why we would advocate for more subjects for this study is for a greater opportunity to sample this heterogeneity. A greater number of subjects might permit identification of a subgroup for whom speech and articulation rate are especially sensitive biomarkers with a strong neural correlate that might not be a universal attribute of all depressed subjects.

In considering the quality of the phoneme rate model's fit to phoneme variability, we must ask fundamentally if we have taken the model as far as it can go in terms of performance given its current state of complexity. All models require a design trade-off between simplicity and complexity. Our model limited complexity to two main rate parameters (plus the articulator parameters that were part of the vocal tract model) that operated at low-level motor control. However, because speech rate control can occur at multiple levels, we may consider adding language production models that include syntax, semantic, and prosodic influences. We also observe that having a single pair of parameters for the entire passage also limits model complexity. Instead, these parameters could come from a probability distribution as a way of incorporating additional variability.

As a complementary approach to the level rate model we investigated based on DIVA, we might also return to studying the GODIVA model and its influence on rate. While GODIVA is primarily concerned with the order of phonemes, it does offer an avenue for influencing the rate of phonemes as well. We hypothesize that we could influence phoneme rate through GODIVA by modulating the difficulty with which successive phonemes are selected and transferred to DIVA for execution. GODIVA to DIVA integration has been studied previously by Civier et al. [30] though it was done with a focus on stuttering and indirect changes in speaking rate rather than phoneme rate variability.

6.3 Extensions

6.3.1 Model Improvements and Useage

This thesis has contributed a systems level approach to understanding phoneme rate as a biomarker of depression. We advanced the use of computational models for their ability to infer hidden parameters of the system, guide future research direction, summarize knowledge, and make quantitative predictions.

We considered models of speech production and neural function, and our model of phoneme rate variability in particular is a starting point for additional investigations into other disorders. We embedded two key parameters within a control frame work that has an articulatory synthesis based plant, an idealized auditory system and forward

model, and a novel means of sequencing phonemes to generate phoneme rate. With the model as a starting platform, one could investigate rate effects under compromise to these modules. For example, auditory damage as a lesion to the auditory system distorts feedback. Do we expect hearing impaired speakers to manifest different changes in phoneme rate? For Parkinson's, disease, we could model changes to the go strength parameter itself. In autism, a general disorder of prediction [134], we could introduce mismatches between the actual and internal models of production.

A question not addressed by this thesis is the homunculus problem. What is the ultimate determiner of speech rate? In other words, this thesis considered low level rate control and suggests additional levels of control built upon low level constraints, but what planner controls the planner? Here we would introduce reward based learning as a driving measure in which environmental exposure creates acoustic targets of typical durations. A learning model will try to emulate these acoustics subject to the physical constraints of the system. Rate may be partially an emergent property of the system. With this model in mind, we can further propose that between depressed and control individuals, average rate may be a coarse discriminating feature. Instead, we might find that individuals draw rate from a statistical distribution of rates. Rather than defining rate as a fixed quantity per person, we should characterize the distribution of rates to identify biases in the distributions between individuals.

This leads to a broader view of measures of rate in speech. This thesis focused on phoneme rate, but rate as a long term spectral change could be considered. The model used here is a foundation in which the phoneme point targets drive the system, but the goodness of fit would not be the mean absolute error on the phoneme durations but the resulting spectral characteristics of the output.

This thesis's macroscale model is a starting point into which components with additional detail may be added. For example, the auditory system is idealized and simplified to perfectly detect formants. A high fidelity auditory system could be added in which degradation of hearing could be simulated in order to predict changes in phoneme rate. Lesions span a spectrum from having diffuse systemic impact to narrow focal disruptions. By adding additional or high fidelity modules to the system, the different effects of diffuse or focal lesions might be simulated. Additional parameters may also allow estimation of different classes of disorders rather than just depression vs control.

In addition to the computational model of phoneme rate, this thesis introduced a computational model of vocal source control. Naturally there is expected mutual coupling between these formant and fundamental frequency systems, at the least because the vocal source imposes mechanical constraints on switching between voiced and unvoiced sounds. By integrating the vocal source as part of the plant, one could investigate how deficits in vocal source control such as vocal chord paresis affect speech rate.

This thesis advocates for the union of modeling and experimentation, and we sketch how this concept can be developed because of our multi-faceted speech collection undertaken as part of the the neuroimaging dataset construction. Looking across the breadth of this corpus we may find further support for basal ganglia differences in depression

during speaking. One of the other task fMRI experiments conducted in our protocol was the "pa-ta-ka" task. The "pa-ta-ka" task involved repeated execution of "pa-ta-ka" at varying speeds. While not naturalistic speech and not as free form in terms of rate control as the emotional sentences, this task may still highlight differences in rate control between the control and depressed populations. Our model would predict different brain activation in the superior temporal gyrus due to dynamic changes in the auditory width parameters. This would be evidence of control by subjects of the precision of their speech as they switch between the three tasks of producing clear, normal, and rapid speech.

■ 6.3.2 Additional Questions

The over arching applied vision for this thesis has been to track individual change over time. Such an assessment would be a longitudinal assessment as opposed to what has been analyzed in this thesis which is cross-sectional assessment. Another name for this problem is to track the state of a person over time, a property that fluctuates, versus a long term characteristic, a person's trait, such as depressed or not depressed. Given this gap, how would we change our system to support tracking within subject changes? Initially such a system could be trained using the cross-sectional data. For a new subject, the cross-sectionally trained classifier could report its unthresholded score, which varies between zero and one. This score provides a finer grained measure for analysis through time. Then, unsupervised techniques for change detection in timeseries could applied to this quantity (see Chandola et al. [26] for a review).

However, we can also take our cue from speech recognition applications in their early days. In a personalized system, we would use intermittent survey responses from the subject or the subject's caregiver to adapt the model to the subject. As a concrete example, we can follow the approaches of Reynolds et al. [110] and Williamson et al. [143] who essentially created a Gaussian mixture model density function for a generic subject and then adapted the means of the Gaussians using subject data.

Within subject tracking of Parkinson's severity has been attempted by Tsanas et al. [130] using speech. However, a review of the example figure in their paper suggests that there is significant variability in the estimate of severity relative to the actual change in severity. In depression, in the AVEC dataset, we informally noticed a similar phenomenon in which patients do not change their severity level much relative to their average over the course of time that they participate in the study. Taking these two cases together, the small within subject changes and the large variability in severity estimates for the two disorders suggests that tracking depression effectively will almost certainly require some form of supervised subject adaptation as opposed to unsupervised change detection.

A practical question that remains to be addressed in both cross-sectional and longitudinal applications is the effect of microphone quality. Presumably, a system would operate on audio data from smart-phone microphones or web cameras. A detailed study of microphone impact across a range of devices would be needed to evaluate impact of

microphones on assessing depression severity. This problem may be especially acute in the longitudinal case in which a person may switch devices from a web cam to a smart-phone to a clinician's office microphone during the course of monitoring.

As we reflect upon the performance of phoneme rate as a biomarker, we might ask what would be the best investment of research money going forward with respect to speech biomarkers? This thesis focused on a narrow acoustic aspect of speech as a demonstration of an integration of speech signal processing, neuroimaging, and computational modeling. However, what else might be fruitfully explored in the speech signal? We highlighted several additional features in the previous chapter. Among these features, phoneme rate itself might give way to a more robust measure of timing such as some measure of spectral change. A spectral change measure could be less susceptible to noise that would cause phoneme dropout or miss-classification of phonemes while still capturing essential elements of low-level acoustic dynamics.

■ 6.3.3 Neuroimaging at Finer Time Scales

Finally, we close this section by returning to the idea of phoneme at the phoneme level, as opposed to articulation rate that is an average phoneme rate. Our fMRI investigation used articulation rate because of the paucity of support for estimating phoneme durations on a single, short sentence. However, as an intermediate step between phoneme level rate and articulation rate, we can aggregate the phonemes into larger clusters to allow greater sample support within a sentence. For example, we can consider consonant rate vs vowel rate at the sentence level, and construct a fMRI task contrast of consonant rate > vowel rate. We do so, and show a result of the contrast in Figure 6.1.

We observe that the precuneus is less active for controls than depressed subjects in the consonant vs vowel rate contrast. Given that abnormal precuneus connectivity is implicated in depression, at least in resting state, a precuneus difference is not unreasonable to expect [47]. Furthermore, the precuneus is involved in visual spatial attention switching [24, 139]. While subjects do not have visual feedback of their articulators, we might speculate that precuneus activity is more generally tied to switching between types of motor activity where consonant motor activity could be viewed as distinctly different than vowels. We stress this is a preliminary investigation in neural differences in depression under a finer temporal analysis of speech than in Chapter 3, but we view this analysis as a natural step towards characterizing phoneme rate.

■ 6.4 Looking to the Future

Why study biomarkers and models of biomarkers of disorders? Ultimately, it is to improve the quality of life of affected individuals and their loved ones. Through biomarkers of a disorder, we can identify problems before they might otherwise become symptomatic, and we can tailor treatment that is specific for that individual. Our current health system nominally attempts to fulfill these aims but falls short both in terms of the treatment timeline and in acknowledging how every single person's disorder is

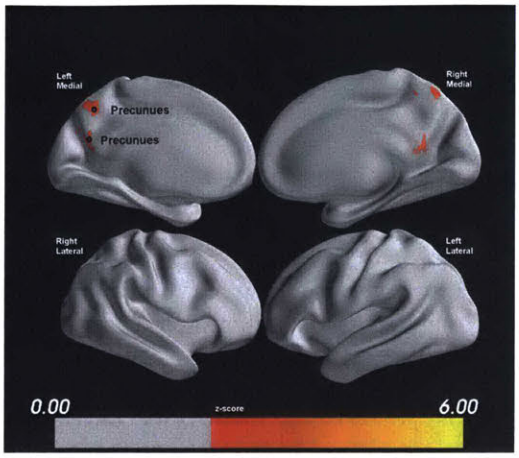


Figure 6.1: There is less activation in the precuneus (both a dorsal segment and ventral segment shown by two black circles) in controls relative to depressed subjects when producing consonants as opposed to vowels. L_1 contrast: consonant rate > vowel rate, Group: controls > depressed. Multiple comparison corrected.

ultimately unique to that person. From a timeline standpoint, we need biomarkers for early indicators of disease, and from a treatment efficacy standpoint we need to quickly establish what is working and what is not and adapt accordingly.

As discussed in Chapter 1, our decision to focus on phoneme rate and depression was a self-imposed limitation that clarified the research program we conducted that brought together analysis of speech, neuroimaging, and psychological assessment. Phoneme rate is just one aspect of speech, and speech itself is just one signal among many observables (e.g., heart rate, accelerometry, sleep habits, social engagement, medical history, and genetics). The smart-phone acts as a common integration point for measuring and processing this information and communicating knowledge to the individual and their care provider. Therefore, we envision our features-from-computational-models-approach transitioning to smart phonemes or household monitors that passively analyze microphone and other sensor signals to monitor health (see "MIT VoiceUp" as a mobile platform supporting this aim).

Depression is one heterogeneous disorder among many including autism, traumatic brain injury, Parkinson's, and schizophrenia. However, our approach generalizes to these other sensing modalities and disorders, and affords an opportunity to view disorders in terms of model and data-driven biotypes ("biologically distinctive phenotypes" [31]) instead of these classical labels. As an example of this approach for depression, Drysdale et al. [39] used resting state fMRI to discover patterns that allowed prediction of treatment responses for different subgroups. Without identification of subgroups, we may miss treatment strategies that would be effective for some individuals because the strategies were not effective for many individuals.

Computational models summarize system level knowledge in a human interpretable form and still capture knowledge precisely by using mathematical descriptions of the system. This facilitates communication among researchers, and by making clear what is considered essential detail now, today's models inform tomorrow's experimental data collections and hypothesis tests. With a model of the system, therapeutic interventions might be devised and tested *in silico*. Finally, models provide a mechanistic framework to perform inference on variables not directly measured by experiments. We have demonstrated phoneme rate and depression as one instance of using computational models through this thesis. We look forward to how integrated, systems level computational modeling will yield additional developments in tracking and treating psychiatric disorders.

Appendix A

The General Linear Model

In this appendix, we review the general linear model (GLM) and illustrate with examples the types of analyses we will conduct. The GLM is a framework that explains the measured brain response within each voxel as a linear combination of experimental factors [70]. These factors could include whether or not the person is speaking and how fast the person is speaking. As we shall see, the GLM is not limited to explaining only the measured hemodynamic response within one subject. It can be used to explain differences in brain responses between groups of subjects using the same math. Because of the complex nature of our experiments, these sections should act as foundation with which to understand our results. Material in this section comes from numerous standard expositions on the general linear model for fMRI analysis including FMRIB Analysis Group and MGH [46], Lindquist and Wagner [82].

■ A.1 GLM Terminology and Significance Tests

The General Linear Model (GLM) is an equation that states that a linear combination of factors, x_1, x_2, x_3, x_n, using coefficients, β_i, can explain an observation, y. The factors x in the GLM may include products of some or all of the factors e.g. $x_1 \cdot x_2, x_1 \cdot x_2 \cdot x_3$. Since the GLM is formulated as a regression problem, it is possible to perform statistical tests on the significance of the β coefficients as well as linear combinations of the β coefficients. Unless noted otherwise, if we say β_i is significant, we mean we have rejected the null hypothesis that $\beta_i = 0$. The procedures for determining significance will be covered later in A.1.2.

Before moving to our example, we also define a contrast. A contrast is a linear combination of some set of β's. Another name for the β_i values are parameter estimates (PE), so another name for a contrast is a contrast of parameter estimates, abbreviated as COPE. A contrast can be tested for significance in the same way that an individual β might be tested for significance.

■ A.1.1 Main Effects, Conditional Effects, and Interactions

To establish a common set of terms for describing our results, we use a general linear model with two factors and the product of the factors. Table A.1 summarizes the language we will use. The terms are "main effect", "conditional effect", and "interaction".

Table A.1: How to describe conclusions from the significance of β values in GLM equations. $\beta \neq 0$ denotes β as statistically significant [57].

Example Equation	β Significance	Statement
$y = \beta_1 x_1 + \beta_2 x_2$	$\beta_i \neq 0$	There exists a main effect of x_i.
$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$	$\beta_1 \neq 0$	There exists a conditional effect of x_1.
$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$	$\beta_3 \neq 0$	There exists an interaction effect between x_1 and x_2.

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 \tag{A.1}$$

If β_3 is not significant, then y may depend only on x_1 or x_2 independent of the other factor. In this reduced model,

$$y = \beta_1 x_1 + \beta_2 x_2. \tag{A.2}$$

In the case of a model which does not include a product of terms, a test on the significance of β_1 or β_2 is a test for the main effect of the corresponding factor, x_1 or x_2. β_1 and β_2 are called main effect coefficients. The interpretation of β_1 and β_2 is straightforward in the reduced model: for every one unit change in x_i value, there is a change of β_i independent of the value of the other factor.

If however there is a significant β_3, then the interpretation of the β_1 and β_2 change, and the terms used to describe them change as well. If β_3 is significant, then β_1 and β_2 are conditional effects instead of main effects of x_1 and x_2. A change in x_1 results in a change in y that depends on the value of x_2. Furthermore, the name given to a β that weights a term that is a product of factors is an interaction effect. β_3 is the interaction effect in this example, and if β_3 is significant, we say there is a significant interaction between x_1 and x_2.

Unfortunately, it can be confusing to change the adjective of "main" for "conditional" depending on the presence or absence of an interaction. Therefore, we will only use "effect" from now on, but we will always say "interaction" or "interaction effect" when we refer to a product of several regressors.

■ A.1.2 GLM Significance Tests

We have just introduced how to "read" the results of significance tests on the β's, and now we overview how to perform the significance tests.

Because the General Linear Model is a multiple linear regression equation, we restate several key facts without proof about the estimated $\hat{\beta}$'s when using maximum

likelihood estimation. Before we give these results, first, we introduce our matrix notation for multiple linear regression. Lower case bold letters denote column vectors. $\boldsymbol{y}$ in particular is a column vector of observations, β is a single coefficient and $\boldsymbol{\beta}$ is the vector of all coefficients. A hat over a variable denotes an estimate quantity as opposed to the true, known value which does not have a hat (e.g., $\hat{\beta}$ vs β.) Uppercase, bold letters denote a matrix such as $\boldsymbol{X}$. $\boldsymbol{X}$ in particular is called the design matrix. $\boldsymbol{x}_i^T$ is a row vector of features from $\boldsymbol{X}$ that are linearly combined by $\boldsymbol{\beta}$ to create the observation y_i.

Using matrix notation, the general linear model can be stated as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{A.3}$$

where $\boldsymbol{\epsilon}$ is a column vector of zero mean errors which we will assume is normally distributed. In the general case, ϵ_i in $\boldsymbol{\epsilon}$ do not need to be independent or even uncorrelated. Therefore, the random vector $\boldsymbol{\epsilon}$ is characterized by its mean, $\mathbf{0}$ and its covariance matrix, $\boldsymbol{V}$. If the errors are uncorrelated, then $\boldsymbol{V}$ has a simple form of $\boldsymbol{V} = \sigma^2\boldsymbol{I}$, where σ^2 is the true variance of the errors and $\boldsymbol{I}$ is the identity matrix.

Our first fact is that the maximum likelihood estimate of the unknown parameters $\boldsymbol{\beta}$ is unbiased[82]. Therefore, the expected value of $\hat{\boldsymbol{\beta}}$ is the true $\boldsymbol{\beta}$[82]:

$$\text{Fact 1: } E\left[\hat{\boldsymbol{\beta}}\right] = \boldsymbol{\beta} \tag{A.4}$$

Our second fact from Lindquist and Wagner [82] is that the variance of the estimated $\hat{\boldsymbol{\beta}}$, denoted as $Var\left(\hat{\boldsymbol{\beta}}\right)$), is:

$$\text{Fact 2: } Var\left(\hat{\boldsymbol{\beta}}\right)) = \left(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}. \tag{A.5}$$

The third fact is that a statistical test on the significance of β or generally the contrast $\boldsymbol{c}^T\boldsymbol{\beta}$ is a t-test with test statistic of

$$t = \frac{\boldsymbol{c}^T\hat{\boldsymbol{\beta}}}{\sqrt{Var\left(\boldsymbol{c}^T\hat{\boldsymbol{\beta}}\right)}} = \frac{\boldsymbol{c}^T\hat{\boldsymbol{\beta}}}{\sqrt{\boldsymbol{c}^T Var\left(\hat{\boldsymbol{\beta}}\right)\boldsymbol{c}}} \tag{A.6}$$

and degrees of freedom equal to $N - p$ where N is the number of observations in $\boldsymbol{y}$ and p is the number of parameters in $\boldsymbol{\beta}$ assuming $\boldsymbol{X}$ is full rank.

The final piece of information needed to apply equation A.6 is $\boldsymbol{V}$. Within a single subject's fMRI run, the errors are not uncorrelated, $\boldsymbol{V} \neq \sigma^2\boldsymbol{I}$, because the hemodynamic response function introduces correlation between sampled time points within a voxel. However, assuming a pre-whitening stage has been applied in which the correlation between time points has been estimated (the details are beyond the scope of this appendix), we can assume $\boldsymbol{V} = \sigma^2\boldsymbol{I}$. Unfortunately, we still do not know the true unknown variance σ^2, and this needs to be estimated from the data. σ^2 is estimated as the sum of the squared errors, $e_i = y_i - \hat{y}_i = y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}$, divided by the number of

degrees of freedom. Written in matrix notation,

$$\hat{\sigma}^2 = \frac{\boldsymbol{e}^T\boldsymbol{e}}{N-p} \tag{A.7}$$

We summarize these statements through a fMRI example for a single brain voxel, of a single fMRI run, for a single subject. The measured brain response at each time point is collected as a column vector $\boldsymbol{y}$. The design matrix, $\boldsymbol{X}$ has columns for motion regressors (the position and rotation of the subject's head as well as the first derivative of these quantities) as well as outlier time points. The design matrix also has the experimental variables which we want to be the factors whose linear combination explains the brain response: these are for example, whether or not the subject is speaking or not, how fast they are speaking, and the sentence type. A linear regression solver computes $\boldsymbol{\beta}$ and the desired contrasts, variances, test statistics, and p values. At this point, the analyst can determine whether or not the contrasts of interest are significantly different from zero and report results using terminology from Table A.1.

■ A.1.3 GLM Across Subjects

At the group level, meaning across subjects, the analysis is similar but not identical. Rather than jointly solving for estimates of $\boldsymbol{\beta}$ across all subjects, it is computationally more tractable to estimate $\boldsymbol{\beta}$ per subject and the the variance per subject for those parameters. Then, the group level analysis can perform another test on the subject level results to determine a final conclusion.

The group analysis GLM for the β associated with a particular regressor, e.g., the β_i associated with phoneme rate, collects the $\hat{\beta}_i$ from each subject into a single "observation" vector $\hat{\boldsymbol{\beta}}_i$. The design matrix is constructed with a mean regressor $\beta_{group,mean}$ (e.g., a column of ones), and the regressor of interest (e.g., depression severity, $\beta_{group,depression}$). Then the GLM solves for the group level $\boldsymbol{\beta}_{group}$, and the group level β for the regressor of interest is tested for statistical significance. Symbolically, we write the group GLM equation as

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{i}} = \boldsymbol{X}_{group}\boldsymbol{\beta}_{group} + \boldsymbol{\epsilon}. \tag{A.8}$$

Referring back to Table A.1, we can read off how to report in words the result of a statistically significant β. In equation A.8, there is no *apparent* product of regression factors, only the mean regressor and the group regressor of interest (e.g. depression severity). Therefore, if the $\hat{\beta}_{group}$ value for depression severity is significant, we say that there is a main effect of depression severity with respect to the subject level regressor whose β was used in the left hand side of the group GLM A.8.

However, it is also correct to say that there is an interaction of the group level regressor with the subject level regressor according to the vocabulary in Table A.1. Though the group level GLM equation does not explicitly show an interaction, we can substitute the group level equation into the subject level equation. We revert back to writing out the GLM terms instead of using matrices for clarity.

Table A.2: Fictional subjects and their fictional brain activations in a single voxel

Subject	BDI	Brain Activity Emotional Sentence	Brain Activity Neutral Sentence
A	20	2	3
B	40	2.5	5
C	60	3.1	7.5

For a simple model with one regressor at the subject level, x_1, and a mean offset, and one regressor at the group level, x_g and a mean offset (aka a mean regressor aka a "one"), we can write the subject level and group level equations,

$$y = \beta_0 + \beta_1 x_1 + \epsilon \tag{A.9}$$
$$\beta_0 = \beta_{00} + \beta_{01} x_g \tag{A.10}$$
$$\beta_1 = \beta_{10} + \beta_{11} x_g, \tag{A.11}$$

we can substitute the group level equation into the subject level equation to get the full model at the subject level

$$y = (\beta_{00} + \beta_{01} x_g) \cdot 1 + (\beta_{10} + \beta_{11} x_g) x_1 + \epsilon \tag{A.12}$$
$$y = \beta_{00} + \beta_{01} x_g + \beta_{10} x_1 + \beta_{11} x_g x_1 + \epsilon. \tag{A.13}$$

Now the brain activity, y, for the subject is explained in terms of the subject level regressor x_1, the group level regressor x_g, and the interaction of the subject level with the group regressor $x_g x_1$.

Using the vocabulary rules in Table A.1, if β_{01} or β_{10} is significant, there is an effect of the group level regressor or the subject level regressor respectively, and if β_{11} is significant, there is an interaction of the group level regressor with the subject level regressor.

We graphically illustrate examples of these terms in the next section.

■ A.2 Interpreting GLM Results

As a simple, fictional case, we pretend we have three subjects each with their own Beck Depression score (BDI). Furthermore, each of these subjects spoke both an emotional sentence and a neutral sentence, and we recorded the brain response of each subject for each sentence at a single voxel within the brain. Table A.2 shows the collected data.

Our subject level regressor is the sentence type and our group level regressor is the BDI score, so the full model for a subject looks like

$$y = \beta_{00} + \beta_{BDI} x_{BDI} + \beta_{emo} x_{emo} + \beta_{emo,BDI} x_{BDI} x_{emo} + \epsilon. \tag{A.14}$$

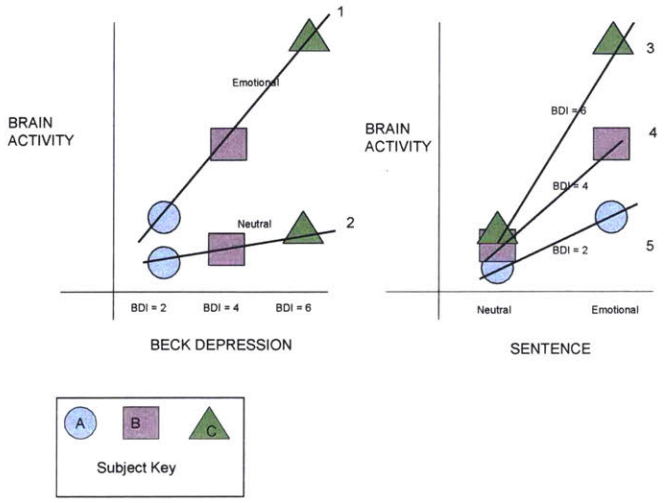


Figure A.1: Two equivalent graphical representations of the fictional data in Table A.2.

■ A.2.1 GLM Analysis Questions

This data can be analyzed several ways to answer questions about whether depression or emotion or their interaction has an effect. Recall that "an effect of x" means that x has a statistically significant contribution to explaining measured brain activity. Mathematically, the β in equation A.14 is non-zero.

However, because we discussed the mathematics already and to build intuition, we will keep these examples at the qualitative level. In essence, the statistical significance test on whether or not the β of interest is 0 is a check for whether or not the data has a zero or non-zero slope with respect to the regressor of interest.

We plot the data from our table in Figure A.1. Note carefully that the same data is represented in both the left and right subplots. However, the x-axis can be changed from the BDI (left) to the emotionality (right).

Question 1: Is there an effect of depression on brain activity?

Let us recall the definition of effect: if depression level affects brain activity, then brain activity depends on depression level keeping other experimental factors such as sentence type constant. The only other experimental factor in this example is the sentence emotionality. Therefore, to answer the question, we can fix the sentence emotionality to be an emotional sentence (line 1 in the left subplot). Then, we look for whether the slope of the line that connects the three subjects in the emotional sentence condition is zero or non-zero. If the slope of the line is zero, then brain activity is independent

of depression severity when speaking emotional sentences. In this example, the slope is non-zero. Consequently, we can conclude that brain activity depends on depression severity when speaking an emotional sentence.

What about when speaking neutral sentences? Again, we check to see if there is a non-zero slope to line 2 in the figure. Once again, there is a non-zero slope, so we can conclude that brain activity depends upon depression level.

We can arrive at these same two conclusions by looking at the right hand subplot instead of the left hand subplot. If we look at the emotional sentence, we can check if depression level affects brain activity by observing whether or not the three data points lie on top of one another or whether they are spread out. We see that that the data points in the emotional sentence category are well spread apart so depression strongly affects brain activity when speaking an emotional sentence. The wide spread of data in the right hand plot for the emotional sentence exactly reflects the large slope in the left hand sub plot.

Similarly, we conclude that depression affects brain activity when speaking neutral sentences because the data points in the right hand subplot do exhibit a small, non random spread, but the spread is smaller than when the sentence is an emotional sentence.

Question 2: Does sentence emotion affect brain activity?

To determine whether sentence emotion affects brain activity, we follow exactly the same procedure as before except we hold depression constant and look at changes of brain activity with respect to changes in emotion. Concretely, we consider the left hand subplot. We fix a depression level, such as BDI = 2 for subject A. Then, we check whether or not subject A's brain activity is different for the two sentences. They are. We check subject B, and again we observe different brain activity depending on the sentence. The same is true for subject C.

As with the previous question, we reach the same conclusion by considering the right subplot. To determine if sentence emotion affects brain activity, we fix a depression level such as BDI = 2 for subject A. Then, we check to see if the slope of the line connecting the two emotion points for subject A is zero or non-zero. If the slope is non-zero, then there is an effect. We see the slope is non-zero which agrees with our conclusion from the left hand subplot. The other two subjects support the same conclusion, sentence emotion does affect brain activity.

Question 3: Does depression interact with sentence emotion?

Up to this point, we considered two basic questions that checked for effects of an experimental variable with brain activity. However, we may wish to ask a sophisticated question: does depression interact with sentence emotion? To answer this question, we use the right hand subplot. If depression interacts with sentence emotion, then the change in brain activity when the sentence emotion changes should depend on depression. We consider the slope of the line for each subject and we compare the

slopes. If the slopes are different, then the change in brain activity for a change in sentence emotionality does depend on depression. If the slopes are the same, in other words, we see parallel lines, then depression does not interact with sentence emotion. We see that the lines are not parallel and therefore conclude that depression does interact with emotion.

As with the two previous questions, we can answer this question by using the left subplot too. If using the left subplot, the slope of line 1 is the change in brain activity with respect to a change in depression for a fixed emotion. If the slope of the line is different for different emotions, then there is a depression by emotion interaction.

Summary

In summary, when testing for effects of experimental manipulation (BDI, sentence emotion) on brain activity, we check for non-zero slopes. When testing for interactions between experimental manipulations, we check for slopes that are different depending on the experimental manipulation. The reader should note that the preceding three questions and analysis hold identically when considering phoneme rate instead of sentence emotion. Finally, all of these qualitative remarks are intended to provide an intuition for how an interaction is determined. The test for an effect or interaction is a rigorous mathematical operation that uses Student's t-test.

■ A.2.2 Phoneme rate, depression, and emotion

Another important topic of investigation is the potential for effects among depression, emotion, and phoneme rate. These effects can be analyzed entirely independent of brain activity, but still we can use the same mathematical frame work. In this case, we can ask all the same questions as before but replace brain activity with phoneme rate.

We show one plot similar to before and we show an alternative plot to provide a different, but equivalent perspective. We also simplify the problem to two subjects, one depressed and one control, or equally valid in interpretation, an average of depressed and an average of control subjects but we distinguish between happy, sad, and neutral sentence types.

Figure A.3 shows phoneme rate frequencies for sentences of different emotion and for control vs depressed subjects. phoneme rate is on the x axis, and the number of times that phoneme rate is recorded is on the y axis. The blue distributions correspond to depressed subject and the red to healthy control. The far left red and blue distributions correspond to the sad condition, the center two distributions correspond to the neutral sentence distribution, and the right most two distributions correspond to the happy sentence condition.

We can ask three questions: does sentence emotionality affect phoneme rate? Does depression affect phoneme rate? Does sentence emotionality interact with depression?

Does sentence emotionality affect phoneme rate? Yes, we see clearly that the bell curves for the control fall at different places on the x axis depending on the sentence emotion. We see the distributions for depressed subjects are similarly spread out.

Equivalently, there is a non-zero slope in the left hand subplot for each subject in Figure A.2.

Does depression affect phoneme rate? Yes, for each fixed sentence emotionality, the distribution for controls is centered to the right of the distribution for depressed subjects. Equivalently, there is a non-zero slope in the right hand plot of Figure A.2.

Does depression interact with sentence emotion? This is an interesting question because in words, it is the same question we posted in the previous subsection. However, in the previous case, we used brain activity as the third unstated variable whereas here we are using phoneme rate. In the earlier case, we checked for differences in the change in brain activity with respect to depression or equivalently changes in brain activity with respect to emotion. Here, we check for differences in the change in phoneme rate with respect to depression among each of the three sentence emotions. In this example, the change in phoneme rate with respect to depression is the same. Each of the three blue curves is shifted the same amount to the left of the red curve for the same sentence type. Equivalently, the lines are parallel in Figure A.2. Consequently, there is no depression interaction with sentence emotion in the acoustic data.

This may seem counter-intuitive, but it is an experimentally possible result. In this situation, where the brain data suggests a depression by emotion interaction but the acoustic data does not, we can hypothesize another unknown mechanism at work. This unknown mechanism is somehow decoupling the speech system acoustic changes associated with depression from the speech system changes associated with sentence emotion.

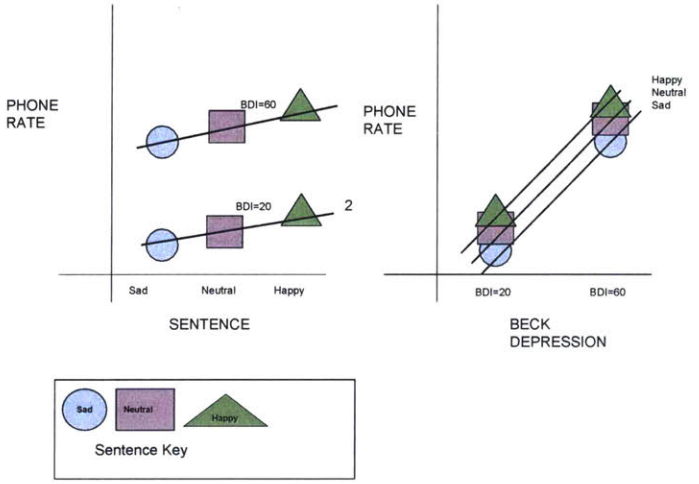


Figure A.2: Mock data demonstrating possible relationships or lack thereof between phoneme rate, depression, and sentence emotion.

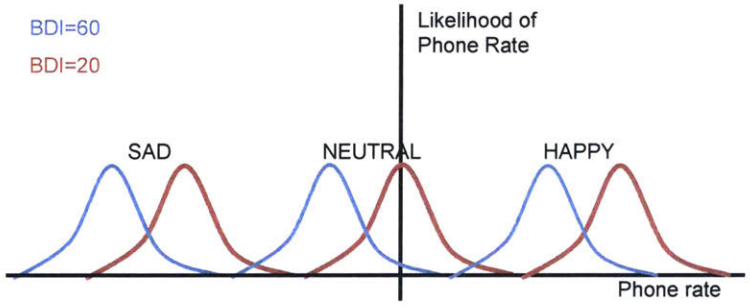


Figure A.3: Hypothesized dependence of phoneme rate, depression, and emotional sentences.

Appendix B

The Complete MRI Protocol

The complete MRI protocol consisted of the following scans and tasks.

Table B.1: Complete scan protocol

Scan	Synopsis	Purpose
Resting	-	Intrinsic connectivity from task-free function
Diffusion	-	Intrinsic connectivity from anatomy
T_1	-	Structural
Pa-ta-ka	Repeat aloud of "pa-ta-ka" at various rates	Articulation effects
Pitched Sentences	Read aloud sentences at various pitches	Prosody
Non-word repetition	Repeat aloud single, made-up words	Novel sequencing of phonemes
Emotional face matching	Match faces with different emotional expressions	Canonical control vs depressed task
Emotional Sentences	Read aloud sentences with an implicit emotional valence	Brings together speech and emotion, and by extension depression
Held Vowels	Pronounce aloud [AH], [IY], and [OH] vowels at low, normal, and high pitches	Vocal source control
Pitched Non-words	Read aloud non-words with "?", "!" and "." emphasis	Investigate linguistic prosody
Movie Trailer	Watch and listen to a movie trailer	Prompt brain responses to a natural stimulus
T_2	-	Structural

Appendix C

Phoneme Duration Correlations

We plot the top 12 phonemes whose mean duration exhibit the largest magnitude Spearman correlation with depression severity. We show all three read passages, and report the correlation (rho) and p value (uncorrected). The error bars are plus and minus one standard error of the mean (sample standard deviation divided by the square root of the number of times the phoneme occurred)[1].

[1] phn_dur_table.ipynb

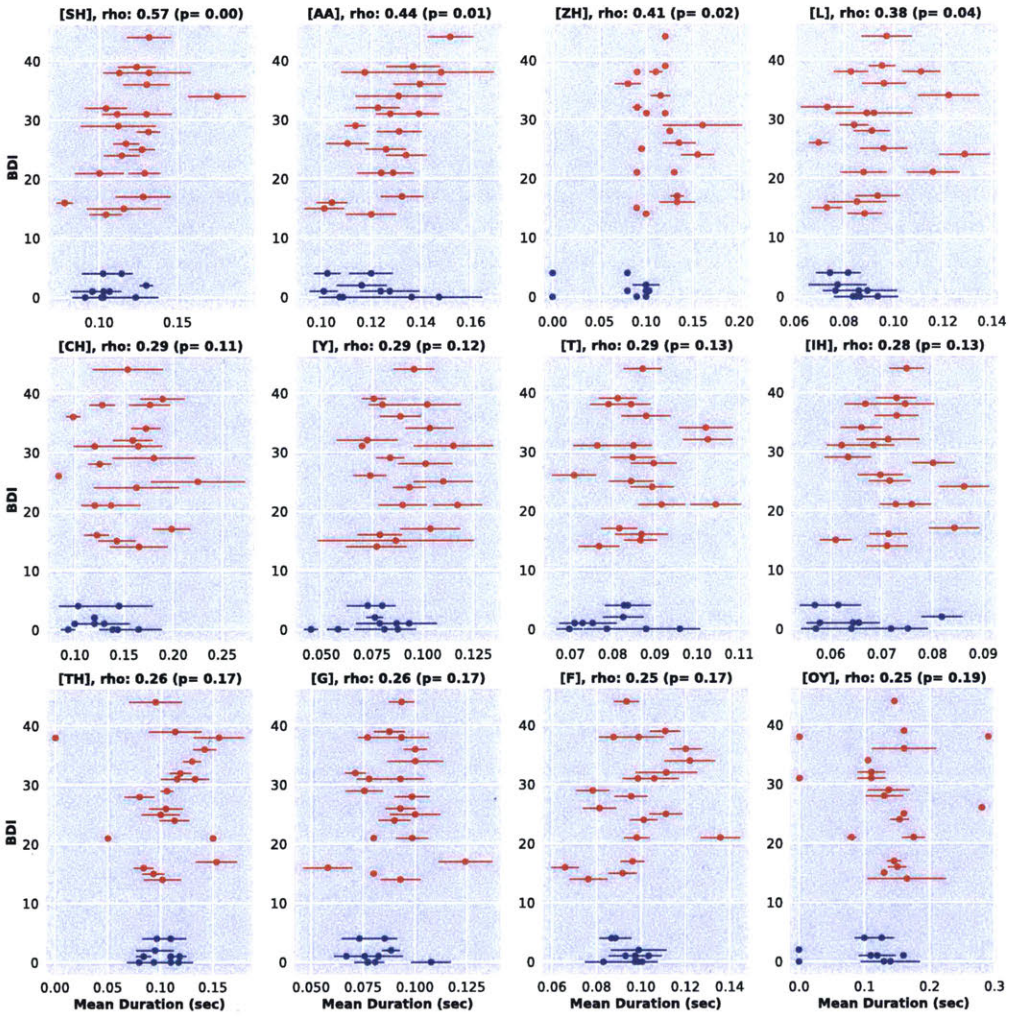


Figure C.1: BDI vs phoneme durations: Rainbow

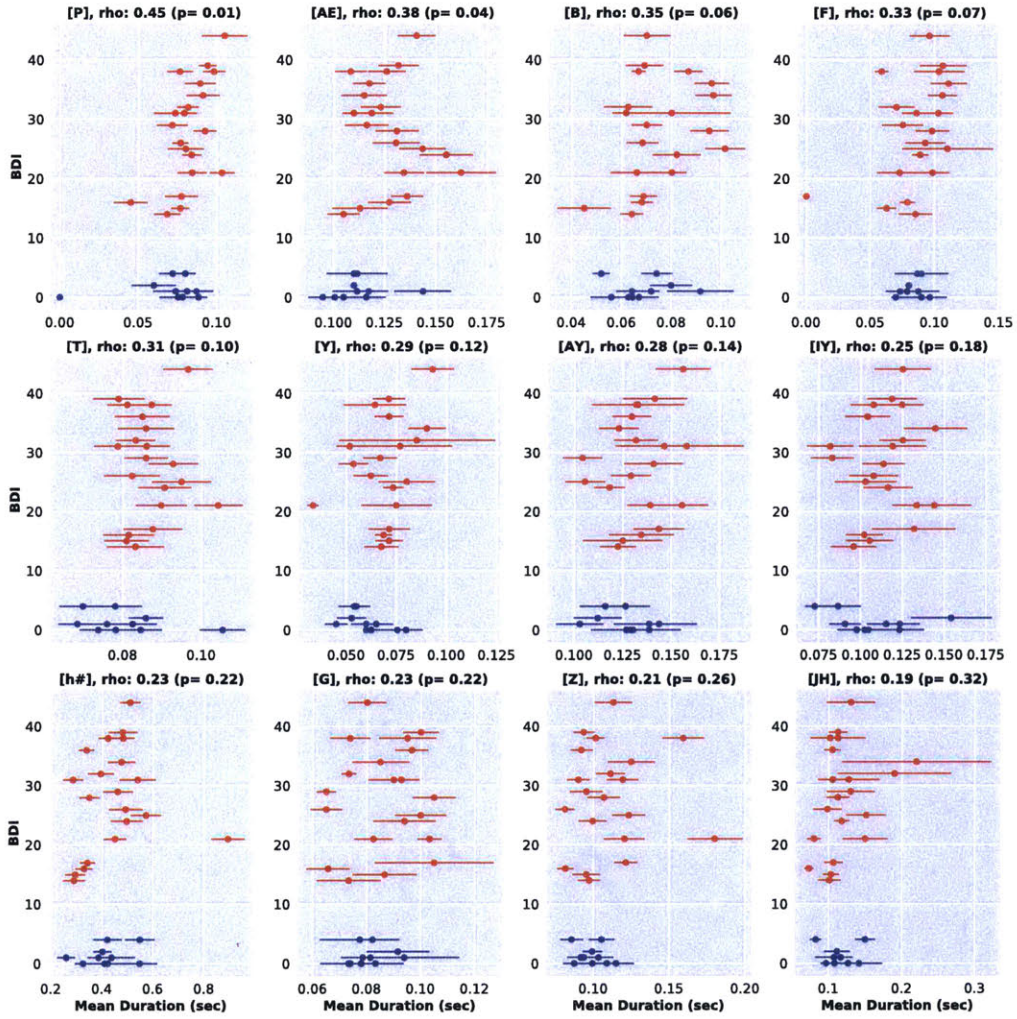


Figure C.2: BDI vs phoneme durations: Caterpillar

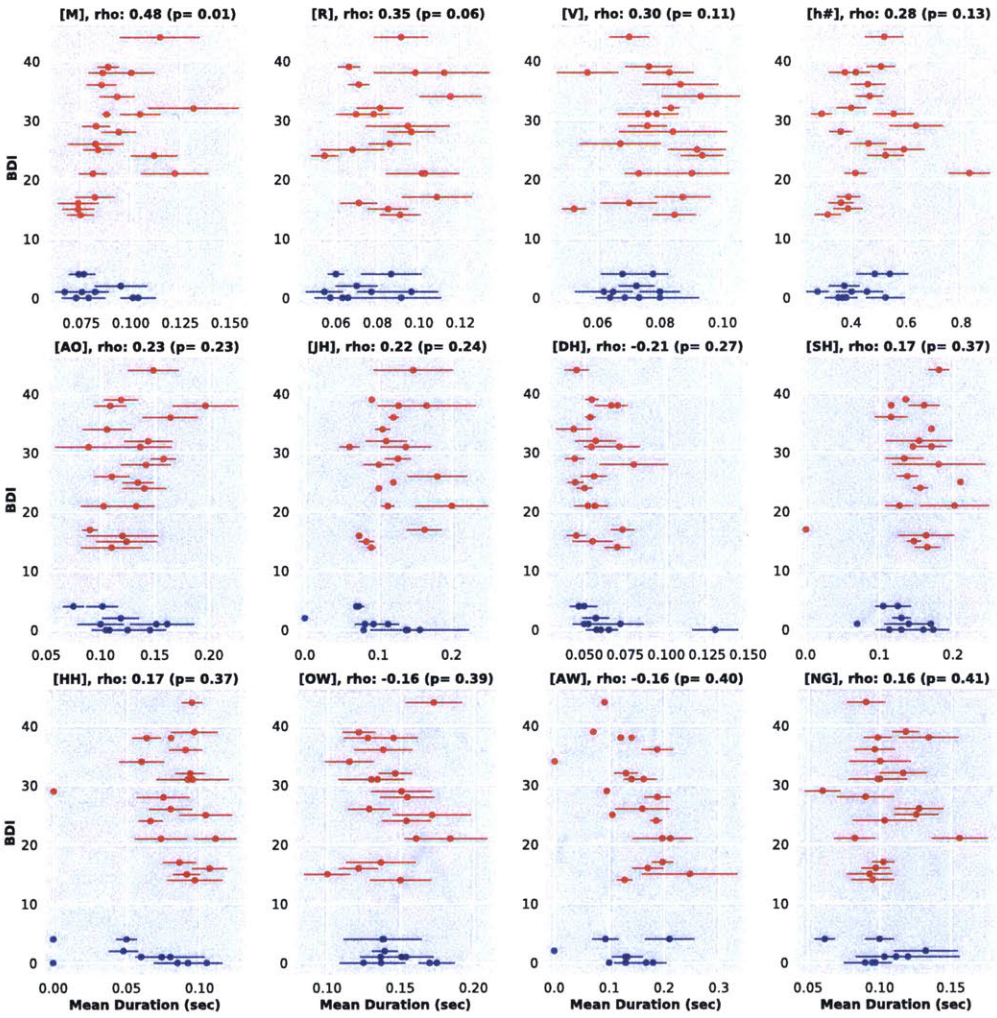


Figure C.3: BDI vs phoneme durations: Grandfather

Bibliography

[1] H. Ackermann, K. Mathiak, and A. Riecker. The contribution of the cerebellum to speech production and speech perception: clinical and functional imaging data. *The cerebellum*, 6(3):202–213, 2007.

[2] G. W. Ahava and C. Iannone. Is the beck depression inventory reliable over time? an evaluation of multiple test-retest reliability in a nonclinical college student sample. *Journal of Personality Assessment*, 70(2):222–231, 1998.

[3] J. Ashburner. *SPM12 Manual.* Functional Imaging Laboratory, Wellcome Trust Centre for Neuroimaging, v6906 edition, 2012.

[4] A. P. Association. *Diagnostic and Statistical Manual of Mental Disorders.* American Psychiatric Publishing, Arlington, VA, 5 edition, 2013.

[5] A. P. Association et al. *Diagnostic and statistical manual of mental disorders (DSM-5®).* American Psychiatric Pub, 2013.

[6] I. P. Association. International phonetics association, Mar 2017. URL `https://www.internationalphoneticassociation.org/`.

[7] B. B. Avants, N. J. Tustison, G. Song, P. A. Cook, A. Klein, and J. C. Gee. A reproducible evaluation of ants similarity metric performance in brain image registration. *Neuroimage*, 54(3):2033–2044, 2011.

[8] N. Bailey, K. Hoy, J. Maller, D. Upton, R. Segrave, B. Fitzgibbon, and P. Fitzgerald. Neural evidence that conscious awareness of errors is reduced in depression following a traumatic brain injury. *Biological psychology*, 106:1–10, 2015.

[9] A. T. Beck, C. H. Ward, M. Mendelson, J. Mock, and J. ERBAUGH. An inventory for measuring depression. *Archives of general psychiatry*, 4(6):561–571, 1961.

[10] A. T. Beck, R. A. Steer, and G. K. Brown. *Manual for the Beck depression inventory-II.* The Psychological Corporation, San Antonio, TX, 1996.

[11] B. M. Ben-David, M. I. Moral, A. K. Namasivayam, H. Erel, and P. H. van Lieshout. Linguistic and emotional-valence characteristics of reading passages for clinical use and research. *Journal of Fluency Disorders*, 49:1–12, 2016.

[12] P. Birkholz. Modeling consonant-vowel coarticulation for articulatory speech synthesis. *PloS one*, 8(4):e60603, 2013.

[13] H. Blumenfeld. *Neuroanatomy through clinical cases*. Sinauer Associates, 2010.

[14] P. Boersma et al. Praat, a system for doing phonetics by computer. *Glot international*, 5(9/10):341–345, 2002.

[15] J. W. Bohland and F. H. Guenther. An fmri investigation of syllable sequence production. *Neuroimage*, 32(2):821–841, 2006.

[16] J. W. Bohland, D. Bullock, and F. H. Guenther. Neural representations and mechanisms for the performance of simple speech sequences. *Journal of cognitive neuroscience*, 22(7):1504–1529, 2010.

[17] S. Borson, J. Scanlan, M. Brush, P. Vitaliano, and A. Dokmak. The mini-cog: a cognitivevital signs measure for dementia screening in multi-lingual elderly. *International journal of geriatric psychiatry*, 15(11):1021–1027, 2000.

[18] S. Borson, J. M. Scanlan, P. Chen, and M. Ganguli. The mini-cog as a screen for dementia: validation in a population-based sample. *Journal of the American Geriatrics Society*, 51(10):1451–1454, 2003.

[19] E. Bresch, J. Nielsen, K. Nayak, and S. Narayanan. Synchronized and noise-robust audio recordings during realtime magnetic resonance imaging scans. *The Journal of the Acoustical Society of America*, 120(4):1791–1794, 2006.

[20] R. Bruffaerts, A. Bonnewyn, and K. Demyttenaere. The epidemiology of depression in belgium. a review and some reflections for the future. *Tijdschrift voor psychiatrie*, 50(10):655–665, 2008.

[21] J. S. Buyukdura, S. M. McClintock, and P. E. Croarkin. Psychomotor retardation in depression: biological underpinnings, measurement, and treatment. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 35(2):395–409, 2011.

[22] S. Cai, S. S. Ghosh, F. H. Guenther, and J. S. Perkell. Focal manipulations of formant trajectories reveal a role of auditory feedback in the online control of both within-syllable and between-syllable speech timing. *Journal of Neuroscience*, 31 (45):16483–16490, 2011.

[23] G. J. Canter. Speech characteristics of patients with parkinson's disease: I. intensity, pitch, and duration. *Journal of Speech & Hearing Disorders*, 1963.

[24] A. E. Cavanna and M. R. Trimble. The precuneus: a review of its functional anatomy and behavioural correlates. *Brain*, 129(3):564–583, 2006.

[25] Center for Behavioral Health Statistics and Quality. Key substance use and mental health indicators in the united states: Results from the 2015 national survey on drug use and health (hhs publication no. sma 16-4984, nsduh series h-51), 2016.

[26] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):15, 2009.

[27] D. K. Chhetri, J. Neubauer, and D. A. Berry. Neuromuscular control of fundamental frequency and glottal posture at phonation onset. *The Journal of the Acoustical Society of America*, 131(2):1401–1412, 2012.

[28] D. G. Childers and J. A. Diaz. Speech processing and synthesis toolboxes, 2000.

[29] G. Ciccarelli, T. Quatieri, and S. Ghosh. Neurophysiological vocal source modeling for biomarkers of disease. In *Interspeech*, pages 1200–1204, 2016.

[30] O. Civier, D. Bullock, L. Max, and F. H. Guenther. Computational modeling of stuttering caused by impairments in a basal ganglia thalamo-cortical circuit involved in syllable selection and initiation. *Brain and language*, 126(3):263–278, 2013.

[31] B. A. Clementz, J. A. Sweeney, J. P. Hamm, E. I. Ivleva, L. E. Ethridge, G. D. Pearlson, M. S. Keshavan, and C. A. Tamminga. Identification of distinct psychosis biotypes using brain-based biomarkers. *American Journal of Psychiatry*, 173(4):373–384, 2015.

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

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

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

[35] F. L. Darley, A. E. Aronson, and J. R. Brown. *Motor speech disorders*. Saunders, 1975.

[36] C. M. Dayan and V. Panicker. Hypothyroidism and depression. *European thyroid journal*, 2(3):168–179, 2013.

[37] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz. Predicting depression via social media. In *ICWSM*, page 2, 2013.

[38] R. S. Desikan, F. Ségonne, B. Fischl, B. T. Quinn, B. C. Dickerson, D. Blacker, R. L. Buckner, A. M. Dale, R. P. Maguire, B. T. Hyman, et al. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *Neuroimage*, 31(3):968–980, 2006.

[39] A. T. Drysdale, L. Grosenick, J. Downar, K. Dunlop, F. Mansouri, Y. Meng, R. N. Fetcho, B. Zebley, D. J. Oathes, A. Etkin, et al. Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nature Medicine*, 2016.

[40] H. Ellgring and K. R. Scherer. Vocal indicators of mood change in depression. *Journal of Nonverbal Behavior*, 20(2):83–110, 1996.

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

[42] G. Fairbanks. *Voice and articulation: Drillbook.* Harper & Brothers, 1940.

[43] G. Fairbanks. Systematic research in experimental phonetics: 1. a theory of the speech mechanism as a servosystem. *Journal of Speech & Hearing Disorders*, 1954.

[44] B. Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012.

[45] P. M. Fitts. The information capacity of the human motor system in controlling the amplitude of movement. *Journal of experimental psychology*, 47(6):381, 1954.

[46] FMRIB Analysis Group and MGH. Fsl course, 2016. URL `https://fsl.fmrib.ox.ac.uk/fslcourse/`.

[47] P. Fossati. Epa-1043–the eye of the self: precuneus and depression. *European Psychiatry*, 29:1, 2014.

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

[49] K. Friston and W. Penny. Post hoc bayesian model selection. *Neuroimage*, 56(4): 2089–2099, 2011.

[50] K. Friston, J. Mattout, N. Trujillo-Barreto, J. Ashburner, and W. Penny. Variational free energy and the laplace approximation. *Neuroimage*, 34(1):220–234, 2007.

[51] K. J. Friston, L. Harrison, and W. Penny. Dynamic causal modelling. *Neuroimage*, 19(4):1273–1302, 2003.

[52] P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Machine learning*, 63(1):3–42, 2006.

[53] S. Ghosh. *Understanding Cortical and Cerebellar Contributions to Speech Production Through Modeling and Functional Imaging.* PhD thesis, Boston University, 2005.

[54] P. Gómez-Vilda, M. Vicente-Torcal, J. M. Ferrández-Vicente, A. Álvarez-Marquina, V. Rodellar-Biarge, V. Nieto-Lluis, and R. Martínez-Olalla. Parkinsons disease monitoring from phonation biomechanics. In *International Work-Conference on the Interplay Between Natural and Artificial Computation*, pages 238–248. Springer, 2015.

[55] K. Gorgolewski, C. D. Burns, C. Madison, D. Clark, Y. O. Halchenko, M. L. Waskom, and S. S. Ghosh. Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Frontiers in neuroinformatics*, 5:13, 2011.

[56] S. Gosh. Vocal tract models. `https://github.com/satra/VocalTractModels`, 2001.

[57] K. Grace-Martin. Testing and dropping interaction terms in regression and anova models. `http://www.theanalysisfactor.com/testing-and-dropping-interaction-terms/`, 2011.

[58] P. E. Greenberg, A.-A. Fournier, T. Sisitsky, C. T. Pike, and R. C. Kessler. The economic burden of adults with major depressive disorder in the united states (2005 and 2010). *J Clin Psychiatry*, 76(2):155–162, 2015.

[59] F. H. Guenther. Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological review*, 102(3):594, 1995.

[60] F. H. Guenther. *Neural Control of Speech.* Mit Press, 2016.

[61] F. H. Guenther, S. S. Ghosh, and J. A. Tourville. Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and language*, 96(3):280–301, 2006.

[62] L. M. Hall, B. Klimes-Dougan, R. H. Hunt, K. M. Thomas, A. Houri, E. Noack, B. A. Mueller, K. O. Lim, and K. R. Cullen. An fmri study of emotional face processing in adolescent major depression. *Journal of affective disorders*, 168: 44–50, 2014.

[63] J. P. Hamilton, A. Etkin, D. J. Furman, M. G. Lemus, R. F. Johnson, and I. H. Gotlib. Functional neuroimaging of major depressive disorder: a meta-analysis and new integration of baseline activation and neural response data. *American Journal of Psychiatry*, 2012.

[64] P. A. Harris, R. Taylor, R. Thielke, J. Payne, N. Gonzalez, and J. G. Conde. Research electronic data capture (redcap)a metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of biomedical informatics*, 42(2):377–381, 2009.

[65] Harvard Center for Morphometric Analysis. Harvard-oxford cortical and subcortical structural atlases, May 2017. URL `https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/Atlases`.

[66] G. Hickok and D. Poeppel. The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8(5):393–402, 2007.

[67] J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler. Acoustic characteristics of american english vowels. *The Journal of the Acoustical society of America*, 97 (5):3099–3111, 1995.

[68] J. F. Houde and S. S. Nagarajan. Speech production as state feedback control. *Frontiers in human neuroscience*, 5:82, 2011.

[69] R. Huerta-Ramírez, J. Bertsch, M. Cabello, M. Roca, J. M. Haro, and J. L. Ayuso-Mateos. Diagnosis delay in first episodes of major depression: a study of primary care patients in spain. *Journal of affective disorders*, 150(3):1247–1250, 2013.

[70] S. A. Huettel, A. W. Song, and G. McCarthy. *Functional magnetic resonance imaging*, volume 1. Sinauer Associates Sunderland, 2004.

[71] P. Indefrey and W. J. Levelt. The spatial and temporal signatures of word production components. *Cognition*, 92(1):101–144, 2004.

[72] B. D. Institute. Bipolar disorder self test, May 2017. URL `https://www.blackdoginstitute.org.au/mental-health-wellbeing/bipolar-disorder/bipolar-disorder-self-test`.

[73] S. D. Kauer, S. C. Reid, A. H. D. Crooke, A. Khor, S. J. C. Hearps, A. F. Jorm, L. Sanci, and G. Patton. Self-monitoring using mobile phones in the early stages of adolescent depression: randomized controlled trial. *Journal of medical Internet research*, 14(3):e67, 2012.

[74] J. P. Kincaid, R. P. Fishburne Jr, R. L. Rogers, and B. S. Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, DTIC Document, 1975.

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

[76] B. J. Kröger, J. Kannampuzha, and C. Neuschaefer-Rube. Towards a neurocomputational model of speech production and perception. *Speech Communication*, 51(9):793–809, 2009.

[77] A. C. Lammert, C. H. Shadle, S. S. Narayanan, and T. F. Quatieri. Investigation of speed-accuracy tradeoffs in speech production using real-time magnetic resonance imaging. *Interspeech 2016*, pages 460–464, 2016.

[78] C. R. Larson, G. B. Kempster, and M. K. Kistler. Changes in voice fundamental frequency following discharge of single motor units in cricothyroid and thyroarytenoid muscles. *Journal of Speech, Language, and Hearing Research*, 30 (4):552–558, 1987.

[79] C. R. Larson, K. W. Altman, H. Liu, and T. C. Hain. Interactions between auditory and somatosensory feedback for voice f 0 control. *Experimental Brain Research*, 187(4):613–621, 2008.

[80] K.-F. Lee and H.-W. Hon. Speaker-independent phone recognition using hidden markov models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(11):1641–1648, 1989.

[81] J. Lenzo. The cmu pronouncing dictionary. `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`, 2017.

[82] M. Lindquist and T. Wagner. Module 22: Inference- contrasts and t-tests. `https://www.coursera.org/learn/functional-mri/home/welcome`, 2015.

[83] K. Livescu. *Feature-Based Pronunciation Modeling for Automatic Speech Recognition.* PhD thesis, Massachusetts Institute of Technology, 2005.

[84] 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*, 58(3):574–586, 2011.

[85] I. S. MacKenzie. Fitts' law as a research and design tool in human-computer interaction. *Human-computer interaction*, 7(1):91–139, 1992.

[86] S. Maeda. Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In *Speech production and speech modelling*, pages 131–149. Springer, 1990.

[87] H. S. Mayberg. Modulating dysfunctional limbic-cortical circuits in depression: towards development of brain-based algorithms for diagnosis and optimised treatment. *British medical bulletin*, 65(1):193–207, 2003.

[88] T. McGlashan, T. Miller, and S. Woods. Prime early psychosis screening test, May 2017. URL http://www.schizophrenia.com/sztest/primeearlypsychosdetails.htm.

[89] T. McGlashan, T. Miller, and S. Woods. Yale university prime screening test, May 2017. URL http://www.schizophrenia.com/sztest/primetest.pdf.

[90] P. R. Montague, R. J. Dolan, K. J. Friston, and P. Dayan. Computational psychiatry. *Trends in cognitive sciences*, 16(1):72–80, 2012.

[91] E. Moore II, M. A. Clements, J. W. Peifer, and L. Weisser. Critical analysis of the impact of glottal features in the classification of clinical depression in speech. *IEEE transactions on biomedical engineering*, 55(1):96–107, 2008.

[92] J. A. Mumford, J.-B. Poline, and R. A. Poldrack. Orthogonalization of regressors in fmri models. *PloS one*, 10(4):e0126255, 2015.

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

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

[95] Å. Nilsonne. Acoustic analysis of speech variables during depression and after improvement. *Acta Psychiatrica Scandinavica*, 76(3):235–245, 1987.

[96] C. A. Niziolek, S. S. Nagarajan, and J. F. Houde. What does motor efference copy represent? evidence from speech production. *Journal of Neuroscience*, 33 (41):16110–16116, 2013.

[97] J. S. Obeid, C. A. McGraw, B. L. Minor, J. G. Conde, R. Pawluk, M. Lin, J. Wang, S. R. Banks, S. A. Hemphill, R. Taylor, et al. Procurement of shared data instruments for research electronic data capture (redcap). *Journal of biomedical informatics*, 46(2):259–265, 2013.

[98] S. Page. Module 22: Inference- contrasts and t-tests. https://www.coursera.org/learn/model-thinking, 2017.

[99] 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*, 22 (1):1–9, 2013.

[100] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[101] J. W. Peirce. Psychopypsychophysics software in python. *Journal of neuroscience methods*, 162(1):8–13, 2007.

[102] W. D. Penny. Comparing dynamic causal models using aic, bic and free energy. *Neuroimage*, 59(1):319–330, 2012.

[103] J. Perkell, M. Matthies, H. Lane, F. Guenther, R. Wilhelms-Tricarico, J. Wozniak, and P. Guiod. Speech motor control: Acoustic goals, saturation effects, auditory feedback and internal models. *Speech communication*, 22(2-3):227–250, 1997.

[104] M. Petrides. *Neuroanatomy of language regions of the human brain.* Academic Press, 2013.

[105] K. L. Phan, T. Wager, S. F. Taylor, and I. Liberzon. Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in pet and fmri. *Neuroimage*, 16(2):331–348, 2002.

[106] S. Pichon and C. A. Kell. Affective and sensorimotor components of emotional prosody generation. *The Journal of Neuroscience*, 33(4):1640–1650, 2013.

[107] C. Poellabauer, N. Yadav, L. Daudet, S. L. Schneider, C. Busso, and P. J. Flynn. Challenges in concussion detection using vocal acoustic biomarkers. *IEEE Access*, 3:1143–1160, 2015.

[108] C. J. Price. A review and synthesis of the first 20 years of pet and fmri studies of heard speech, spoken language and reading. *Neuroimage*, 62(2):816–847, 2012.

[109] D. A. Regier, W. E. Narrow, D. E. Clarke, H. C. Kraemer, S. J. Kuramoto, E. A. Kuhl, and D. J. Kupfer. Dsm-5 field trials in the united states and canada, part ii: test-retest reliability of selected categorical diagnoses. *American journal of psychiatry*, 170(1):59–70, 2013.

[110] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital signal processing*, 10(1-3):19–41, 2000.

[111] A. Riecker, K. Mathiak, D. Wildgruber, M. Erb, I. Hertrich, W. Grodd, and H. Ackermann. fmri reveals two distinct cerebral networks subserving speech motor control. *Neurology*, 64(4):700–706, 2005.

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

[113] J. B. Russ, R. C. Gur, and W. B. Bilker. Validation of affective and neutral sentence content for prosodic testing. *Behavior research methods*, 40(4):935–939, 2008.

[114] S. Scherer, G. M. Lucas, J. Gratch, A. S. Rizzo, and L.-P. Morency. Self-reported symptoms of depression and ptsd are associated with reduced vowel space in screening interviews. *IEEE Transactions on Affective Computing*, 7(1):59–73, 2016.

[115] H. S. Schroder, T. P. Moran, Z. P. Infantolino, and J. S. Moser. The relationship between depressive symptoms and error monitoring during response switching. *Cognitive, Affective, & Behavioral Neuroscience*, 13(4):790–802, 2013.

[116] K. Setsompop, B. A. Gagoski, J. R. Polimeni, T. Witzel, V. J. Wedeen, and L. L. Wald. Blipped-controlled aliasing in parallel imaging for simultaneous multislice echo planar imaging with reduced g-factor penalty. *Magnetic Resonance in Medicine*, 67(5):1210–1224, 2012.

[117] L. M. Seversky, S. Davis, and M. Berger. On time-series topological data analysis: New data and opportunities. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 59–67, 2016.

[118] W. Shen, C. M. White, and T. J. Hazen. A comparison of query-by-example methods for spoken term detection. Technical report, DTIC Document, 2009.

[119] S. Skodda. Effect of deep brain stimulation on speech performance in parkinson's disease. *Parkinsons Disease*, 2012, 2012.

[120] S. Skodda and U. Schlegel. Speech rate and rhythm in parkinson's disease. *Movement Disorders*, 23(7):985–992, 2008.

[121] O. J. Smith. A controller to overcome dead time. *ISA journal*, 6(2):28–33, 1959.

[122] S. M. Smith, M. Jenkinson, M. W. Woolrich, C. F. Beckmann, T. E. Behrens, H. Johansen-Berg, P. R. Bannister, M. De Luca, I. Drobnjak, D. E. Flitney, et al. Advances in functional and structural mr image analysis and implementation as fsl. *Neuroimage*, 23:S208–S219, 2004.

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

[124] K. N. Stevens. Toward a model for lexical access based on acoustic landmarks and distinctive features. *The Journal of the Acoustical Society of America*, 111 (4):1872–1891, 2002.

[125] B. H. Story and I. R. Titze. Voice simulation with a body-cover model of the vocal folds. *The Journal of the Acoustical Society of America*, 97(2):1249–1260, 1995.

[126] I. R. Titze and B. H. Story. Rules for controlling low-dimensional vocal fold models with muscle activation. *The Journal of the Acoustical Society of America*, 112(3):1064–1076, 2002.

[127] I. R. Titze, E. S. Luschei, and M. Hirano. Role of the thyroarytenoid muscle in regulation of fundamental frequency. *Journal of Voice*, 3(3):213–224, 1989.

[128] J. A. Tourville and F. H. Guenther. The diva model: A neural theory of speech acquisition and production. *Language and cognitive processes*, 26(7):952–981, 2011.

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

[130] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig. Accurate telemonitoring of parkinson's disease progression by noninvasive speech tests. *IEEE transactions on Biomedical Engineering*, 57(4):884–893, 2010.

[131] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig. Novel speech signal processing algorithms for high-accuracy classification of parkinson's disease. *Biomedical Engineering, IEEE Transactions on*, 59(5):1264–1271, 2012.

[132] D. Unabridged. phone, Mar 2017. URL `http://www.dictionary.com/browse/phone`.

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

[134] S. Van de Cruys, K. Evers, R. Van der Hallen, L. Van Eylen, B. Boets, L. de Wit, and J. Wagemans. Precise minds in uncertain worlds: predictive coding in autism. *Psychological review*, 121(4):649, 2014.

[135] A. J. van der Kouwe, T. Benner, D. H. Salat, and B. Fischl. Brain morphometry with multiecho mprage. *Neuroimage*, 40(2):559–569, 2008.

[136] V. van Doren. The smith predictor: A process engineer's crystal ball, May 1996. URL http://www.controleng.com/single-article/the-smith-predictor-a-process-engineer-s-crystal-ball/1e02746896161e157f0a7175f67d37db.html.

[137] Van Essen laboratory. Conte69 human surface-based atlas and reference data. http://brainvis.wustl.edu/wiki/index.php/Caret:Atlases:Conte69_Atlas, 2012.

[138] C. Van Riper. *Speech correction.* Prentice Hall, 1963.

[139] N. Wenderoth, F. Debaere, S. Sunaert, and S. P. Swinnen. The role of anterior cingulate cortex and precuneus in the coordination of motor behaviour. *European Journal of Neuroscience*, 22(1):235–246, 2005.

[140] T. V. Wiecki, J. Poland, and M. J. Frank. Model-based cognitive neuroscience approaches to computational psychiatry clustering and classification. *Clinical Psychological Science*, 3(3):378–399, 2015.

[141] D. Wildgruber, H. Ackermann, and W. Grodd. Differential contributions of motor cortex, basal ganglia, and cerebellum to speech motor control: effects of syllable repetition rate evaluated by fmri. *Neuroimage*, 13(1):101–109, 2001.

[142] J. R. Williamson, D. W. Bliss, D. W. Browne, and J. T. Narayanan. Seizure prediction using eeg spatiotemporal correlation structure. *Epilepsy & Behavior*, 25(2):230–238, 2012.

[143] J. R. Williamson, T. F. Quatieri, B. S. Helfer, R. Horwitz, B. Yu, and D. D. Mehta. Vocal biomarkers of depression based on motor incoordination. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 41–48. ACM, 2013.

[144] J. R. Williamson, T. F. Quatieri, B. S. Helfer, G. Ciccarelli, and D. D. Mehta. Vocal and facial biomarkers of depression based on motor incoordination and timing. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 65–72. ACM, 2014.

[145] J. R. Williamson, T. F. Quatieri, B. S. Helfer, J. Perricone, S. S. Ghosh, G. Ciccarelli, and D. D. Mehta. Segment-dependent dynamics in predicting parkinsons disease. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[146] D. M. Wolpert, Z. Ghahramani, and M. I. Jordan. An internal model for sensorimotor integration. *Science*, 269(5232):1880, 1995.

[147] A. Wright. Homeostasis and higher brain function. In J. H. Byrne, editor, *Neuroscience Online*, chapter 6: Limbic System: Amygdala. The University of Texas Health Science Center at Houston, 2015. Accessed: 2015-04-20.

[148] A. Yau and S. Verma. Laryngeal nerve anatomy. `http://emedicine.medscape.com/article/1923100-overview`, 2013. Accessed: 2016-03-28.

[149] M. Zañartu Salas. *Influence of acoustic loading on the flow-induced oscillations of single mass models of the human larynx.* PhD thesis, Purdue University West Lafayette, 2006.