

5-2021

Quantifying Dimensions of the Vowel Space in Patients with Schizophrenia and Controls

Elizabeth Maneval
William & Mary

Follow this and additional works at: <https://scholarworks.wm.edu/honorsthesis>



Part of the [Phonetics and Phonology Commons](#)

Recommended Citation

Maneval, Elizabeth, "Quantifying Dimensions of the Vowel Space in Patients with Schizophrenia and Controls" (2021). *Undergraduate Honors Theses*. William & Mary. Paper 1660.
<https://scholarworks.wm.edu/honorsthesis/1660>

This Honors Thesis -- Open Access is brought to you for free and open access by the Theses, Dissertations, & Master Projects at W&M ScholarWorks. It has been accepted for inclusion in Undergraduate Honors Theses by an authorized administrator of W&M ScholarWorks. For more information, please contact scholarworks@wm.edu.

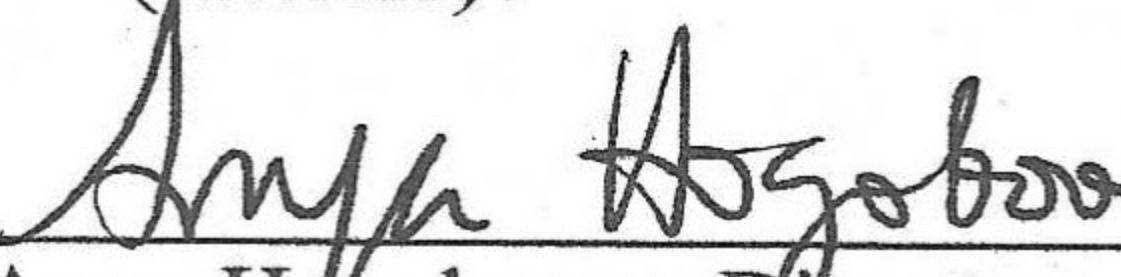
Quantifying Dimensions of the Vowel Space in Patients with Schizophrenia and Controls


A thesis submitted in partial fulfillment of the requirement
for the degree of Bachelor of Science in Linguistics from
William & Mary

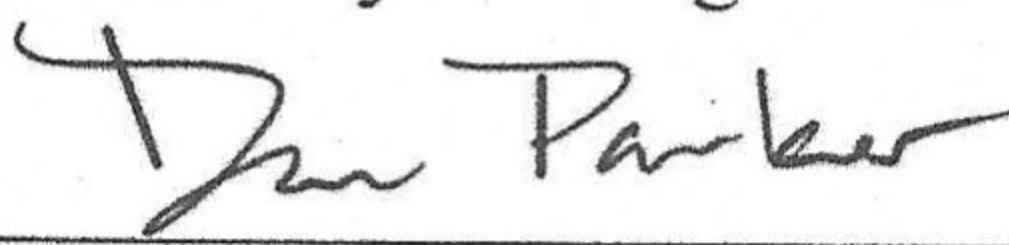
by


Elizabeth Charlotte Maneval

Accepted for Honors
(Honors)


Dr. Anya Hogoboom, Director


Dr. Kaitlyn Harrigan


Dr. Daniel Parker


Dr. Zach Conrad

Williamsburg, VA
May 1, 2021

Acknowledgements

I wish to express my sincerest gratitude to my advisor, Anya Hogoboom, for her patience, insight, and unwavering support during every step of this project. Without her constant encouragement, I never would have completed this thesis. Thanks to Kate Harrigan for her assistance in conceptualizing approaches for analyzing the data. Thanks to Jacob Adams for his tireless work on troubleshooting with the coding involved in FAVE to make this project a reality. Thanks to the William & Mary Linguistics Research Group for their supportive and constructive feedback and ideas as this project progressed. Additional thanks to my family and friends for their grounding support and input on the various iterations of this project.

Abstract

The speech of patients with schizophrenia has been characterized as being aprosodic, or lacking pitch variation. Recent research on linguistic aspects of schizophrenia has looked at the vowel space to determine if there is some correlation between acoustic aspects of speech and patient status (Compton et al. 2018). Additional research by Hogoboom et al. (submitted) noted that measurements of Euclidean distance (ED), which is the average distance from the center of the vowel space to all vowels produced, and vowel density, which is the proportion of vowels clustered together in the center of the vowel space, were significantly correlated for patients with schizophrenia, but not for controls; this correlation was primarily due to a subset of 13 patients. In addition, they found that ED independently was a weak predictor of patient status, but that both density and ED, when used together, were predictors of patient status. This previous study utilized Prosogram (Mertens 2014), a tool that relies on acoustics to sift through the sound files and identify the vowels, which showed unstable reliability in detecting vowels.

Therefore, this research aims to reassess the relationship between the vowel space and patient status by gathering more reliable measurements of the vowels from Hogoboom's dataset by using the forced aligner FAVE (Rosenfelder et al. 2014). We seek to determine if there is a stronger correlation between vowel space usage and patient status than previously found—one that was previously masked by incomplete vowel measurements. Our current research finds that ED is a strong predictor of patient status ($p < 0.05$). While Hogoboom's previous work found that ED and density were independently significant, current work finds that those two variables are correlated. These results show that there is a relationship between ED and an individual's patient status, where patients have lower average ED and controls have higher average ED. Overall, this research clarifies differences in utilization of the vowel space between patients with schizophrenia and controls, which could ultimately be used to create more quantitatively-defined linguistic measurements for diagnosis that are less subject to individual clinical listeners.

1. Introduction

Schizophrenia is a mental disorder with a wide variety of symptoms, divided into positive and negative categories. Negative symptoms are ones where an individual is more withdrawn, which can be a challenging dimension to measure. Current diagnostic tools are prone to racial bias influences by clinicians, prompting a need for more robust and standardized ways to assess speech. In order to do this, this study focuses on analyzing aspects of speech in an attempt to uncover any measurable differences in the speech of patients with schizophrenia and controls.

We seek to build on previous work in finding more quantifiable means for assessing the speech associated with negative symptoms of schizophrenia.

1.1 Schizophrenia

Schizophrenia is a mental disorder that is difficult to diagnose and affects approximately 1% of the population. The *DSM-V* contains six different criteria for diagnosing schizophrenia: characteristic symptoms, social and/or occupational dysfunction, duration of at least six months, schizoaffective and major mood disorder exclusion, substance and/or general mood condition exclusion, and relationship to global development delay or autism spectrum disorder (Tandon et al. 2013). Within the category of characteristic symptoms, there are positive symptoms, like delusions, hallucinations, and disorganized speech, and negative symptoms, including diminished emotional expression, avolition (lack of motivation), and aprosody. Negative symptoms are linked to poorer social outcomes than positive symptoms, but few treatments currently exist to address these debilitating aspects of schizophrenia (Murphy et al. 2006).

Currently, negative symptoms of schizophrenia are rated on the *Scale for the Assessment of Negative Symptoms* (SANS), *Clinical Assessment Interview for Negative Symptoms* (CAINS), and *Positive and Negative Symptom Scale* (PANSS), among others; the rating a patient receives for each is based on an individual clinician's impressions of the patient's speech, rather than a measurable element of speech, and is a culmination of many different aspects of how the individual interacted with the clinician at the time of diagnosis. Some research by Tahir et al. (2019) has found significant correlations between nonverbal speech cues and psychological scale ratings, specifically regarding speech gaps, response times, and mutual silence, among others. While nonverbal cues can provide useful insights into this disorder, verbal cues share the same potential, with the capacity to provide even more information about schizophrenia and its characteristics.

1.2 Linguistic Implications of Schizophrenia

Schizophrenia could have underlying linguistic patterns that have yet to be actively assessed and accounted for in the diagnostic process. Some linguistic elements of speech have the potential to be assessed within the psychological scales currently used to screen for schizophrenia. Within a schizophrenia diagnosis, aprosody is a negative symptom described as a lack of pitch variation in speech. Linguistic approaches have previously informed schizophrenia investigation, including work done by Covington et al. (2012), who found that pitch was not correlated with negative symptom severity. As a follow-up to Covington et al.'s findings Compton et al. (2018), whose data this current study works with a subset of, assessed the acoustic attributes of aprosody in schizophrenia by recruiting patients with schizophrenia, approximately 25% of whom were rated as having aprosody, and controls to complete five tasks: describing pictures for a defined length of time, spontaneously responding to two different

questions, and reading two separate texts aloud. They sought to clarify the acoustic attributes of aprosody in schizophrenia, specifically regarding pitch variation, F1 and F2 measurements which correlate with tongue and mouth movements involved in the pronunciation of vowels, and intensity. Ultimately, they found a significant difference in pitch between patients with aprosody and controls in one of the reading tasks, as well as lower intensity measurements amongst patients with aprosody. Building on that work, Hogoboom et al. (submitted) assessed vowel space reduction in patients with schizophrenia, finding a small subset of patients with a significantly reduced vowel space.

Technological advancements and automated speech assessment tools have been previously used to assess speech in relation to a wide variety of psychiatric disorders, including anxiety, schizophrenia, and PTSD (Low et al. 2020); for schizophrenia, lower speech rate, higher pause duration, and inconsistent findings regarding pitch variation are speech attributes associated with alogia, poverty of speech, and negative symptoms that could be expanded upon to both solidify linguistic assessments and clarify acoustic attributes of the speech of patients with schizophrenia. Hinzen and Rosselló (2015) used speech for analysis in the case of positive symptomatology in schizophrenia, creating a model that combines the three primary positive symptoms and relies on speech as insight into the neurocognitive aspects of a patient's diagnosis; this work focuses on looking at grammar breakdown and a loss of referential capacity as a way to gain neurocognitive insights, rather than relying on the concept of delusional thought. Together, these findings show promising results for relying on more concrete speech-based approaches in diagnosing schizophrenia.

1.3 Negative Symptom Psychological Scales

Currently, negative symptoms are measured through a variety of scales, including the SANS, CAINS, PANSS, *Negative Symptoms Assessment 16* (NSA-16), and *Brief Negative Symptom Scale* (BNSS) (Kumari et al. 2017). The present study utilizes data from patients who were rated on the SANS, CAINS, and PANSS scales. SANS assesses negative symptoms across 25 items on a six-point scale within five broader categories. These five broader categories are affective flattening or blunting, alogia, avolition/apathy, anhedonia/asociality, and attention (Kumari et al. 2017). CAINS assesses negative symptoms through two scales: the Motivation and Pleasure (MAP) Scale, comprised of nine items, and the Expression (EXP) Scale, comprised of four items. These items cover the five categories of blunted affect, alogia, avolition, anhedonia, and asociality (Strauss & Gold 2016). PANSS assesses both positive and negative symptoms, breaking down into three scales: positive, negative, and general psychopathology. The positive and negative scales both have seven items each, whereas the general psychopathology scale has 16 items (Kay et al. 1987). The specific SANS, CAINS, and PANSS items that potentially evaluate measurable phonetic characteristics are SANS 7 (Lack of Vocal Inflections) and 8 (Global Rating of Affective Flattening), CAINS 11 (Vocal Expression), and

PANSS N1 (Blunted Affect). The specific descriptions of each of these items can be found in Appendix A.

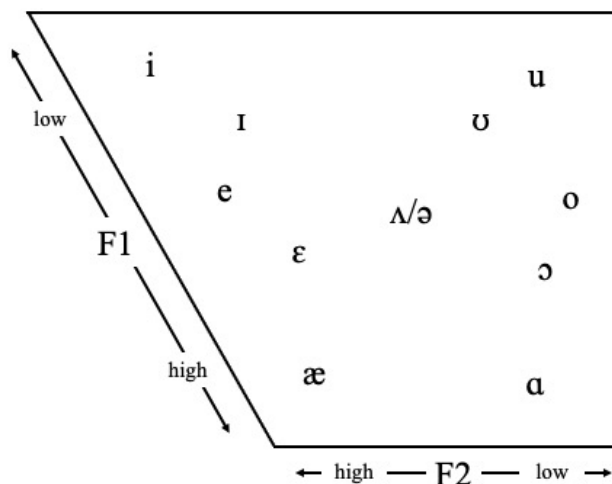
The terms “blunted affect” and “affective flattening” seem to be used somewhat interchangeably; however, upon closer inspection of the criteria for these within the three scales of interest, it can be seen that these two items (SANS 8 and PANSS N1) are not measured with the same overall approach. PANSS N1 relies primarily on facial expressions and gestures, whereas SANS 8 is a broader, more cumulative assessment of the affective flattening or blunting category, with a specific focus on unresponsiveness, inappropriateness, and emotional intensity. While PANSS N1 does not seem to rely on any acoustic criteria, it was included in this analysis in case there were underlying acoustic influences impacting the ratings.

These items (SANS 7, SANS 8, CAINS 11, and PANSS N1) are the ones closest to representing phonetic details, such as vowel space reduction examined in the current study; it is important to note that none of these assessments claims to be able to capture these differences. Rather, if a real difference is found and is therefore perceptible by clinicians, then these are the closest items in which those noticeable attributes would appear.

1.4 Vowel Space

The vowel space is used within linguistics to map where different vowel qualities are produced based on tongue position. F1 and F2 are formants of speech, measured in Hertz, that correspond with tongue height and tongue backness, respectively, and are inherent characteristics of vowels¹. These measures are used to differentiate between different types of vowels. Using inverted F1 and F2 values, Figure 1 illustrates the canonical vowel space of English speakers.

Figure 1. Canonical English vowel space.



Each person utilizes their own breadth of the vowel space. It has been noted that females, who have shorter vocal tracts on average, tend to use broader vowel spaces than males, meaning

¹ F3 is a vowel characteristic also pertinent in languages that have a rounding distinction; English lacks this contrast.

that their vowels are more spread out than those of males (Yang & Whalen 2015). For example, the current study found the following approximate vowel spaces: a male F1 range from 350 to 700 Hz and a female F1 range from 400 to 800 Hz, as well as a male F2 range from 900 to 2100 Hz and a female F2 range of 1000-2500 Hz. Regardless of gender, we see that speakers' F2 values for vowels have a larger range than F1 measurements.

Previous work has noted that there are differences in where vowels are produced within the vowel space across various dialects. For example, Clopper and Pierrehumbert (2008) found that northern US English speakers had a more dispersed vowel space than southern US English speakers. We also see that certain contexts promote phonetic reduction for speakers of Northern and Midland American English; in contexts with high predictability and high frequency of certain words, some vowels were more likely to shift to a more centralized position in the vowel space (Clopper et al. 2016). Other researchers have done work comparing Wisconsin, North Carolina, and Ohio dialects, finding that Wisconsin speakers exhibited features of the Northern Cities Vowel Shift, whereas North Carolina speakers produced more fronted vowels and some monophthongization of the diphthong /aɪ/ (Fox & Jacewicz 2009). We therefore see several factors already known to affect the vowel space.

2. Data Source

The data for this research comes from the previously described Compton et al. (2018) study, which is part of a larger cohort of studies within a National Institute of Mental Health (NIMH) grant that aims to apply computational linguistics to aspects of schizophrenia. We used the data from Task 4, which had the patients read an excerpt from Johanna Spyri's children's story *Heidi* (1998 [1880]); the text of the excerpt can be found in Appendix B. By specifically using the reading task, we could ensure that we would get comparable frequencies of vowels from both patients and controls, something that had the potential to be unevenly distributed in the recordings of tasks with spontaneous speech. Compton et al.'s work focused on using the recordings to provide insights into the acoustic underpinnings of aprosody, whereas this study focuses on mapping and comparing the vowel spaces of patients and controls.

2.1 Participants

There were initially 130 total participants from the Washington D.C. and New York City areas, categorized on binary gender. One participant was excluded from the study because their recording did not reach the halfway point of the reading. Of the remaining participants, 68 were controls and 61 were patients with schizophrenia. A breakdown of the demographic information of the participants is shown in Table 1 below.

Table 1. Demographic information of participants by patient status, race, ethnicity, and gender.

Patient Status	Race	Ethnicity	Gender		Total	
			Male	Female		
Patient	American Indian/Alaska Native	Hispanic or Latinx	0	0	0	
		Not Hispanic or Latinx	0	1	1	
	Asian	Hispanic or Latinx	0	0	0	
		Not Hispanic or Latinx	0	1	1	
	Black/African American	Hispanic or Latinx	1	0	1	
		Not Hispanic or Latinx	33	10	43	
	Native Hawaiian/Other Pacific Islander	Hispanic or Latinx	0	0	0	
		Not Hispanic or Latinx	0	0	0	
	White	Hispanic or Latinx	0	0	0	
		Not Hispanic or Latinx	7	2	9	
	Other	Hispanic or Latinx	2	1	3	
		Not Hispanic or Latinx	2	1	3	
	Total			45	16	61
	Control	American Indian/Alaska Native	Hispanic or Latinx	0	0	0
Not Hispanic or Latinx			1	0	1	
Asian		Hispanic or Latinx	1	0	1	
		Not Hispanic or Latinx	0	0	0	
Black/African American		Hispanic or Latinx	2	0	2	
		Not Hispanic or Latinx	26	16	42	
Native Hawaiian/Other Pacific Islander		Hispanic or Latinx	0	0	0	
		Not Hispanic or Latinx	1	0	1	
White		Hispanic or Latinx	0	3	3	
		Not Hispanic or Latinx	7	6	13	
Other		Hispanic or Latinx	0	3	3	
		Not Hispanic or Latinx	2	0	2	
Total				40	28	68

2.2 Previous Analysis of Data

Previous work by Hogoboom et al. (submitted) focused on vowel space reduction as an indicator of negative symptoms of schizophrenia. This research used Compton et al. (2018)'s data that focused on the acoustic underpinnings of schizophrenia's negative symptoms; Hogoboom et al.'s work specifically used both natural and read speech by the participants as sources of vowels for analysis, with the goal of comparing vowel space sizes between patients and controls. Researchers utilized Euclidean distance (ED) and density as tools for comparing

the vowel spaces. Euclidean distance is a previously well-established tool for measuring the vowel space in linguistics, primarily in assessing vowel mergers and differences (Nycz & Hall-Lew 2013). Euclidean distance, measured in Hertz, takes the average F1 and F2 values for an individual and then measures out to each individual vowel. Essentially, this uses the Pythagorean Theorem by creating a right triangle between the central point of the vowel space and each individual vowel, where that distance is the length of the hypotenuse. The culmination of distances from that average center of their vowel space to each vowel they produced is then averaged to create that individual's average ED. A higher ED represents a more spread out vowel space, whereas a lower ED represents a smaller vowel space. The equation for Euclidean distance for an individual participant's vowels is shown below:

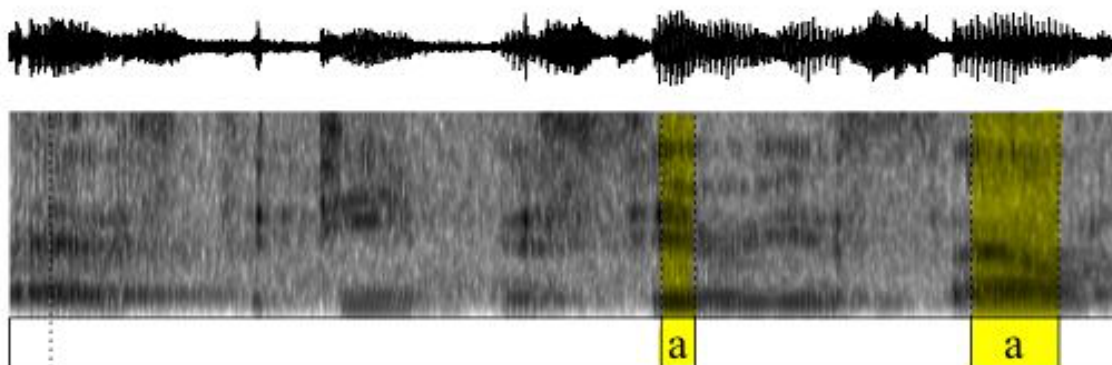
$$\text{Euclidean Distance} = \sqrt{(F1_{\text{participant_mean}} - F1)^2 + (F2_{\text{participant_mean}} - F2)^2}$$

Density, on the other hand, takes one standard deviation around each participant's mean F1 and one standard deviation around their F2 to essentially create a box that contains the vowels of interest. The proportion of vowels within that box is referred to as someone's density. These two measures capture different aspects of the vowel space: Euclidean distance collapses over F1 and F2, while density preserves the differences in F1 and F2 ranges, as we tend to see a larger range of F2 variability in comparison to F1. A higher density correlates with a more reduced and centralized vowel space, whereas a lower density correlates with a more spread out vowel space.

Hogoboom et al.'s work found that density and Euclidean distance were independent predictors of patient status, particularly due to a subset of 13 patients with a substantially reduced vowel space; the vowel spaces of these 13 individuals had low ED and high density. Once those 13 patients were removed from the analysis, ED and density were no longer predictors of patient status. It is possible that these measures of the vowel space were not found to be more robust predictors of patient status due to how the vowels were initially gathered, prior to ED and density measures being taken.

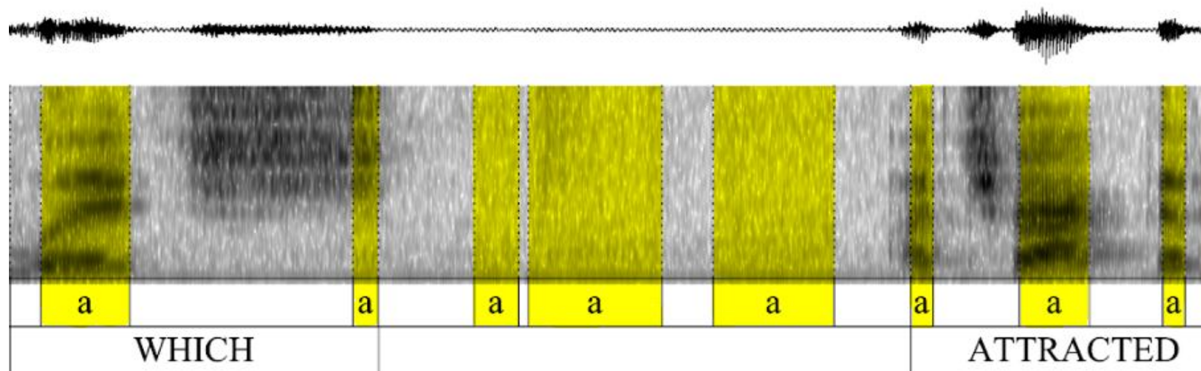
Hogoboom et. al. utilized Prosogram (Mertens 2014) to locate and extract all of the vowels in each sound file for measurements. This program cannot differentiate between vowels, so it labels all vowels it finds with the letter 'a'. Looking more closely at the outputs Prosogram produced, it is evident that the program was not able to find all of the vowels. Figure 2 shows an example excerpt of Prosogram's output for one participant saying the phrase "as the deep mysterious sounds." That specific phrase contains eight vowels, but Prosogram only found two.

Figure 2. Prosogram's output, underidentifying vowels present in the data.



Not only did Prosogram have difficulties finding all the vowels, but it also misidentified various instances where there most certainly was not a vowel, as shown in Figure 3. The excerpt shown is “which attracted” from the first sentence of the reading, where the speaker paused for approximately 740 milliseconds between the two words; Prosogram not only mislabeled part of the fricative as a vowel, but also incorrectly “found” three vowels within that one pause that were not present. This is not to say Prosogram could not locate vowels, as we see it successfully locating all of the vowels in the word “attracted”; rather, it performs well in some instances, but fails at too high of a rate to instill confidence in its reliability.

Figure 3. Prosogram's output, overidentifying vowels not present in the data.



This program primarily relies on acoustics to determine what is a vowel, so it does not have as much support for ensuring accuracy as other programs, like forced aligners, which reference a transcript of the sound file, could provide. Therefore, the relationship between Euclidean distance, density, and patient status that this research found could have been masked by the fact that Prosogram did not have the most reliable collection of vowels prior to analysis.

3. Methods

Our study takes the recorded readings of a passage from a book which served as the fourth recorded task, and first reading task, in Compton et al. (2018) and is also a subset of the recorded data used in Hogoboom et al. (submitted), but uses FAVE (Rosenfelder et al. 2014), a forced aligner which takes both the acoustics and a transcription of each sound file to match each sound in the audio to sounds in the transcript. Using FAVE allows us to gather a more comprehensive set of vowel measurements prior to analysis. This work seeks to collect more reliable measurements of the vowels than Prosogram was able to do, in order to more accurately compare the vowel spaces of patients with schizophrenia to those of controls.

3.1 Forced Aligners

Forced aligners are a newer development within linguistic research that are able to time-align transcriptions with audio files. Forced aligners, like Kaldi, FAVE, and DARLA, among many others, have been used in a variety of contexts within sociophonetics and sociolinguistics, including comparing the vowels of speakers of different dialects and reducing the amount of time sociolinguists spend transcribing interviews. Forced aligners can be semi-automated, requiring researchers to transcribe the speech for the forced aligner to reference, while others are fully-automated, utilizing automatic speech recognition (ASR) tools, similar to those used for Siri and Google Voice. Fully-automated forced aligners can go through to identify and align speech without referencing a transcript; this eliminates the hours of tedious transcription by individual researchers. Some forced aligners are pretrained, like FAVE and DARLA, while others, like Kaldi, require training before they can be deployed for use on a dataset.

Kaldi (Povey et al. 2011) is a fully-automated forced aligner that has to be trained to recognize an individual's speech based on various audio file inputs prior to being able to be used as a forced aligner. This allows the program to become more familiar with a particular person's speech in order to ensure more accurate alignments. Once it has been trained, Kaldi is able to filter through datasets and align speech without requiring a transcript.

Forced Aligner and Vowel Extraction (FAVE), the semi-automated forced aligner utilized in this study, is a program created by Rosenfelder et al. (2014) that combines Evanini et al.'s tool for automatically measuring formants (2009) with University of Pennsylvania's Penn Aligner (Yuan & Liberman 2008). In order to successfully time-align the audio with the transcriptions, FAVE uses the Hidden Markov Model Toolkit (HTK), which has information about the acoustics of speech necessary for the task of forced alignment (Bailey 2016). Upon referencing these dependencies, FAVE is able to discern where to label the start and end of each sound within each word.

Dartmouth Linguistic Automation (DARLA) is a fully-automated forced aligner. When Reddy and Stanford (2015) conducted a study comparing FAVE to DARLA, they found that while DARLA was not always reliable in its transcriptions, both programs produced fairly

comparable formant measurements outputs. DARLA, however, did not represent the Southern Vowel Shift present in the data of some of the speakers as much as FAVE (Reddy & Stanford 2015).

Overall, forced aligners provide researchers with tools that are less time-intensive for transcribing audio files. The advancement within fully-automated forced aligners provides exciting potential for this process to become more reliable and readily deployable in the near future, thus decreasing the time-intensive aspect of forced alignment even more. This study uses FAVE in order to ensure more accurate transcriptions for clarifying the actual vowel measurement differences that may be present in the data; while a fully-automated program would be more time-efficient, it was less feasible for ensuring the close accuracy necessary for clarifying the results of previous research.

3.2 Software

Three pieces of software were used to sift through the data and prepare it for vowel extraction: Praat, ELAN, and FAVE. Praat is a freely-available phonetic software that can edit sound files and pull out measurements from delineated sound files (Boersma & Weenink 2019). ELAN 6.0 is a free annotation tool that can be used to generate time-aligned transcriptions of sound files (ELAN 2020). FAVE (Forced Alignment and Vowel Extraction) v.1.2.2 is a free forced aligner that generates a Praat TextGrid (TextGrids delineate the acoustic signal, as shown by the vertical lines in Figures 4 and 5, below for example) by matching each time-aligned transcription with the corresponding sound file, after referencing the software's editable dictionary (Rosenfelder et al. 2014). This program has the capacity to work through hundreds of files at once.

Figures 4 and 5 revisit the same Prosogram excerpts from before, but with FAVE's more accurate output shown beneath it. These FAVE outputs are adjusted to indicate the actual IPA equivalents, rather than the coding outputs to make the programs easier to compare. Figure 4 clearly shows that FAVE was able to locate all eight vowels in the phrase "as the deep mysterious sounds," while simultaneously differentiating between vowel qualities, in comparison to Prosogram's output of two undifferentiated vowels.

Figure 4. FAVE's comparative output to Prosogram's underidentification of vowels.

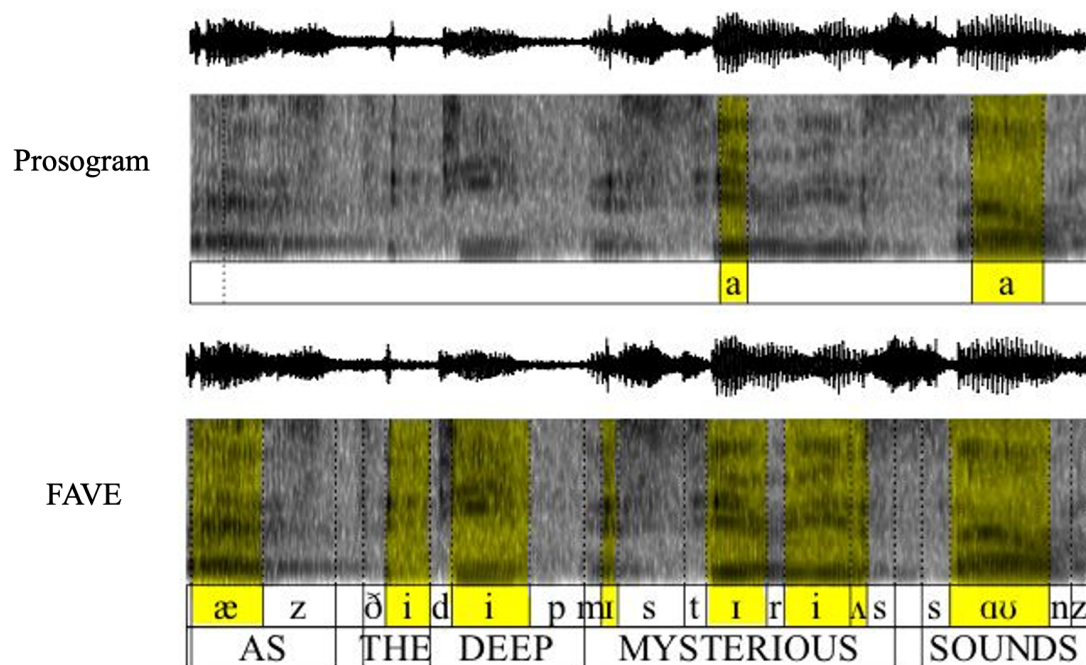
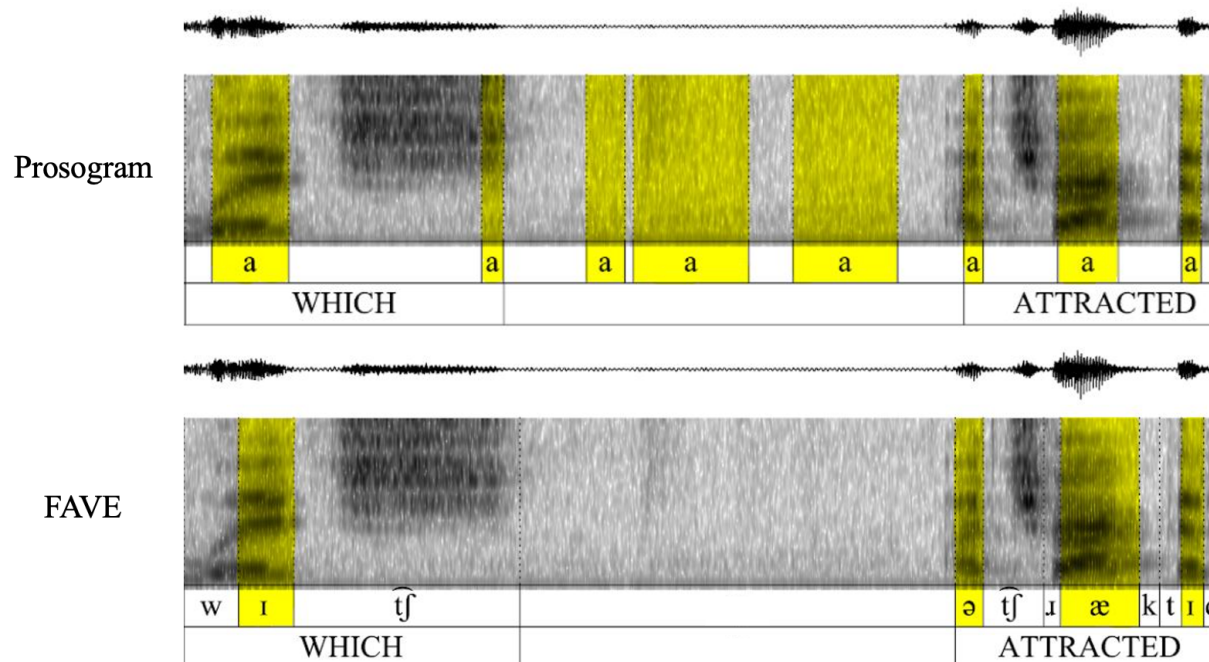


Figure 5 shows how FAVE did not make the same misidentification errors that Prosogram made, both by not denoting any vowels in the silence between words and by refraining from labeling a vowel within the fricative.

Figure 5. FAVE's comparative output to Prosogram's overidentification of vowels.



Overall, these examples show that FAVE can identify vowels more thoroughly and reliably, boosting confidence in utilizing this program for this analysis.

3.3 Workflow

Each sound file from Task 4 in Compton et al. (2018) was edited in Praat to remove the voices that were not those of the participants, as well as other sounds from the surrounding environment, such as loud noises and phones ringing, that could interfere with the forced alignment process later. The trimmed sound file was then opened into ELAN, where it was transcribed. Each intonational phrase was divided into its own chunk in the delineation to ensure better accuracy for the forced aligner. The transcript was exported as a tab-delimited text file and added, with the corresponding trimmed sound file, to the data folder within FAVE.

FAVE had two primary commands for the user to deploy: `checkFAVE` and `runFAVE`. The `checkFAVE` command referenced FAVE's dictionary prior to running the forced alignment. This dictionary utilized a phonetic alphabet, called the ARPAbet, to define each word. This dictionary made distinctions between vowels based on stress: unstressed (coded as '0'), primary stress (coded as '1'), and secondary stress (coded as '2'). The `checkFAVE` command went through each transcription to ensure that every word found in the transcript had a phonetic definition that was compatible with FAVE's dictionary; any words not found in the dictionary were added to the "unknown" text output that FAVE generated for each individual transcript. The user added each unknown word with its own phonetic entry to the dictionary. Once all the words were defined, the `runFAVE` command, which initiated the forced alignment process, was run. This process aligned the transcript with the sound file to generate a Praat TextGrid output. The alignment process for any files that encountered errors was terminated for the user to go back, fix, and run again. The TextGrids were then put back into Praat with their corresponding sound files in order to utilize a Praat script (Lennes 2003) to extract the formant measurements for each sound.

3.4 Dialectal Considerations

Forced aligners are made to accommodate certain language varieties. Minority dialects are consistently unaccounted for in automated transcription tools, requiring researchers to spend time editing the outputs or modifying the dictionaries, like in the present study, to account for variability in speakers' dialects. FAVE, as a pretrained forced aligner, has an editable dictionary that is preprogrammed with words and their corresponding phonetic transcriptions based on Mainstream American English (MAE), which prompts the need for considerations of various phonological patterns within different dialects that could surface in the speech of participants. Regardless of the forced aligner used, it is necessary for researchers to take into consideration different dialectal patterns present in participants' speech during narrow transcription. These considerations allow for more accurate representations of an individual's utterances, which ultimately provides more reliability within the analysis of their speech.

In this reading task, there were some pronunciation differences that required adjustments to the original Heidi excerpt. For example, speakers would encounter the word ‘cupboard’ [kʌbɔ:d], a term which not everyone may have been familiar with, and break it down into ‘cup’ [kʌp] and ‘board’ [bɔ:d]. Additionally, the word ‘shone’ [ʃoʊn] was sometimes pronounced as ‘shun’ [ʃʊn], the word ‘fir’ [fɪr] was said as ‘fire’ [fɪr], and ‘bowed’ [baʊd] was pronounced as [boʊd]. These differences in pronunciation required adjustments to the original transcription of the *Heidi* excerpt that varied by participant, but other phonological differences were more pervasive and consistent with features of the African American English dialect.

3.4.1 African American English

African American English (AAE) has many distinctive features, commonly separating into syntactic and phonological features (Lippi-Green 2011) (Craig et al. 2003). Examples of syntactic features of AAE include the zero copula (MAE: “They’re my favorite”; AAE: “They my favorite”), habitual ‘be’ (MAE: “She’s always tired”; AAE: “She be tired”), and possessive deletion (MAE: “Mary’s book”; AAE: “Mary book”). At the phonological level, speakers of AAE can substitute [d] for [ð] prevocally, where a word like ‘them’, which would be transcribed as [ðem] in MAE, is transcribed as [dem] in AAE. Additionally, the velar [ŋ] becomes the alveolar [n], often word-finally, as seen with a word like ‘stopping’ where MAE produces [stɑpɪŋ] and AAE produces [stɑpɪn]. There is also consonant cluster movement within AAE, where consonants in a cluster can metathesize, like the word ‘ask’ as [æsk] in MAE being [æks] in AAE. There are many more phonological features specific to this dialect which were accounted for in the transcriptions of these sound files when necessary.

Not every speaker of AAE uses every feature of AAE at every point in time; rather, usage of these features varies based on context. Because the data analyzed in this study was based on the reading of a passage, the syntactic features of AAE did not arise; in contrast, the phonological features were widely present. These features were the most relevant for adjusting FAVE’s dictionary to ensure that these phonological patterns of the dialect were being properly accounted for at the transcription level, especially since almost 70% of participants were Black or African American. While not everyone used these features, they frequently surfaced and were important to properly account for in FAVE’s dictionary to ensure more accurate alignments that were phonetically representative of the sounds produced by the participants.

3.5 Dictionary Considerations

While FAVE has been shown to require more labor-intensive edits to transcription for varieties other than Mainstream American English (MAE), it does have the ability to be modified to better represent different dialects, as seen in two studies that assessed its accuracy with varieties of British English (MacKenzie & Turton 2020) (Bailey 2016) and another two studies that used it on Trinidadian English (Meer 2020) (Meer et al. 2021). Bailey (2016) expanded FAVE’s dictionary to account for three phonetic sociolinguistic features of British English and

found that FAVE was able to identify and align sounds with remarkable accuracy, similar to that of human transcribers. While this is great news for this forced aligner, errors still do occur.

MacKenzie & Turton (2020) and Bailey (2016) both found that FAVE's alignment accuracy declines when the speech is faster paced, which is not a decrease in accuracy paralleled by human transcribers. In the Trinidadian English study, researchers assessed the reliability of FAVE and two other forced aligners. FAVE was found to perform fairly well in producing accurate alignments, but still introduced more erroneous alignments of vowels specific to Trinidadian English because it did not have acoustic models to reference within its dependencies (Meer 2020). As can be expected, when the MAE version of FAVE (US-FAVE) was compared to a FAVE program designed with Trinidadian English as a baseline (TRINI-FAVE), TRINI-FAVE more accurately measured the Trinidadian vowels than US-FAVE (Meer et al. 2021).

FAVE has the capacity to work with disfluent speech, such as partial words and word restarts, but we found that attempts to utilize this aspect of FAVE resulted in failed alignments. Therefore, we adjusted our approach to edit the sound files to remove instances of disfluent speech prior to the transcription phase of the workflow. Incomplete words that lacked a vowel were removed because they could not easily be made into a new word in FAVE's dictionary. To ensure that a participant's data was not unnecessarily removed, instances where pretend words could be made to match the recorded speech were capitalized upon. For example, one participant said 'snow' but stuttered on the 'sn' at first--that stutter was removed. Meanwhile, another participant said 'snowsh' [snouʃ] in place of 'snow' [snou], and while that may not be a word, it had a vowel and was added to the FAVE dictionary to avoid editing any participant's data more than necessary and preserve as many vowels as possible.

Dictionary adjustments were crucial beyond simple disfluencies in order to account for pronunciation differences. For example, the name 'Heidi' [hɑɪdi] had various ways of being pronounced, all of which needed to be added to the dictionary as separate entries. Some of these new entries included 'Haydee' [heidi], 'Heldee' [heldi], and 'Heedee' [hidi]. In addition, FAVE had a couple phonetic definitions available for the word 'the': [ðʌ] and [ði]. It relied on information deeper within the dependencies of its coding to determine which transcription to apply to each instance, however it was found to err on the side of assuming the high front tense vowel. In order to avoid misidentification by FAVE, we capitalized on the editable dictionary, using 'thee' [ði] and adding 'thuh' [ðʌ] for times when the version of 'the' was audibly clear at the time of transcription.

Other changes to the dictionary were more predictable, as they aligned with features of African American English (AAE). As mentioned earlier in Section 3.4, speakers of AAE can substitute [d] for [ð] prevocally. This affected the adjustments to the transcription of 'the' that were previously made to include 'dee' [di] and 'duh' [dʌ] when those variations arose. This was present in other words, such as 'although', which appears as [alðou] in MAE and [aldou] in AAE. We encountered interdental fricatives becoming labiodental, like in the word 'underneath', which surfaces as [ʌndəniθ] in MAE and [ʌndənɪf] in AAE. Furthermore, we saw the velar nasal becoming alveolar, as seen in the word 'beginning', which is transcribed as [bɪɡɪnɪŋ] in

MAE and as [bɪɡɪnɪn] in AAE. While this is not an exhaustive list of all the dictionary edits made, it provides a taste of the variation encountered and accounted for in the transcriptions.

3.6 Dataset Creation

The midpoint of each sound's F1 and F2 measurements were extracted with a Praat script (Lennes 2003). This was done in batches divided by gender to account for differences in the number of expected formants (5 within the first 5000 Hz for male voices and within the first 5500 Hz for female voices). The resulting dataset was then trimmed down to only include monophthongs. English's inherent diphthongs [eɪ] and [oʊ] were included due to their minimal change in location between sounds. All consonants and other diphthongs were set aside. Vowels that were outliers were removed by z-scoring the F1 and F2 measurements individually by subject; instances above 3.29 standard deviations and below -3.29 standard deviations of the mean were removed, as they were most likely mismeasurements.

While FAVE makes distinctions between vowels based on three stress levels, there were only approximately seven instances per person of each vowel that carried secondary stress; in order to simplify the analysis, these instances of secondary stress were merged with the primary stress category, thus allowing us to compare stressed and unstressed vowels. Any vowels that did not occur at least 500 times in total were excluded from analysis.

4. Replicating Hogoboom et al.

This research replicates applicable aspects of the Hogoboom et al. (submitted) study, prior to expanding into other analyses. The present study only used data from one reading-based task, whereas Hogoboom et al. utilized data from both reading and spontaneous speech tasks from Compton et al. (2018). This therefore reduced the number of participants from 148 (78 controls, 70 patients) in Hogoboom et al.'s work to 129 (68 controls, 61 patients) in the present study. While Hogoboom et al. did work with both measuring the vowel space and analyzing pitch, this study only looks at replicating the vowel space aspect, as that is the element of the prior study that this one sought to improve upon; pitch was not assessed.

4.1 Results

Table 2 shows the ranges and average measures of Euclidean distance and density by patient status. On average, patients had smaller Euclidean distances and higher densities than controls.

Table 2. Average Euclidean distance and density measurements by patient status.

	Dimension	N	Minimum	Maximum	Mean (\bar{x})	Std. Deviation (s)
Patients	Euclidean Distance (Hz)	61	207.77	436.36	321.8332	52.91738
	Density	61	0.08	0.29	0.1493	0.03750
Controls	Euclidean Distance (Hz)	68	235.18	529.65	350.6907	62.28137
	Density	68	0.07	0.23	0.1390	0.02936

Hogoboom et al. found that ED and density were independently statistically significant predictors of patient status. In determining whether this new data maintained this relationship between ED and density, and that both variables could be put together in a model, we ran a nonparametric correlation and found that the two variables were correlated, as shown in Table 3.

Table 3. Nonparametric correlation between Euclidean distance and density.

Euclidean Distance	Density
Correlation Coefficient	-0.495**
Significance (2-tailed)	<0.001
N	129

We therefore found that, unlike Hogoboom et al., modeling Euclidean distance with density was not predictive; for the remainder of this analysis, we used Euclidean distance, as it was a previously established way to measure the vowel space.

To determine how Euclidean distance, a continuous variable, and patient status, a categorical variable, related to each other, we ran a generalized linear model. As hypothesized, we found a statistically significant impact of gender and patient status on Euclidean distance. There was not a significant interaction term between patient status and gender. This model is shown in Table 4.

Table 4. Generalized linear regression model effects on Euclidean distance by patient status and gender.

Euclidean Distance	Wald χ^2	<i>df</i>	Significance
(Intercept)	4781.353	1	<0.001
Gender	22.504	1	<0.001
Patient Status	7.820	1	0.005
Gender * Patient Status	3.311	1	0.069

We ran a Spearman's rho two-tailed bivariate correlation on Euclidean distance and the aprosody ratings and found that there was not a significant correlation between ED and any of the psychological scale items discussed in Section 1.3, as shown in Table 5.

Table 5. Spearman's rho correlation of Euclidean distance and aprosody ratings.

Euclidean Distance	SANS 7	SANS 8	CAINS 11	PANSS N1
Correlation Coefficient	0.006	0.008	-0.002	0.89
Significance (2-tailed)	0.965	0.953	0.989	0.501
N	60	60	61	60

These SANS, CAINS, and PANSS ratings were only available for patients, as the controls did not come into the study with any ratings on these scales. Note that CAINS 11 has 61 total ratings because one patient was only rated on that scale. While Hogoboom et al.'s work did not look at any correlations to PANSS, the present study included PANSS N1 (Blunted Affect) because it fit with the other measures; we still found no correlation between the PANSS score and Euclidean distance.

4.2 Replication Discussion

Results from this study both reflected and clarified results from Hogoboom et al.'s work. We found a significant relationship between Euclidean distance and patient status ($p=0.005$); in doing so, we clarified the importance of different measures of the vowel space, finding that density did not prove to be as useful for capturing vowel space differences in our dataset as it was in their work. Additionally, we reaffirmed that there was no correlation between Euclidean distance and negative symptom psychological scale ratings.

4.2.1 Vowel Space

While the previous work of Hogoboom et al. found that Euclidean distance and density were both predictors of patient status when used together ($p=0.012$ and $p=0.018$, respectively), the current work finds that those two variables are correlated. Density was a measurement Hogoboom et al. created to better capture the vowel space distribution; this correlation we found between Euclidean distance and density indicates that density is not providing us with additional information not already captured by Euclidean distance. While density preserves the differences in variability in F1 and F2 ranges, it is not a difference we see the motivation for within this dataset; rather, the differences we find can be adequately accounted for by using only Euclidean distance. Therefore, we elected to continue the analysis with a focus on Euclidean distance, as that is the measurement already established in the literature. Our results found that Euclidean distance is a significant predictor of patient status. These results show that there is a relationship between Euclidean distance and an individual's status as a patient or control, where patients have a lower overall Euclidean distance and controls have a higher Euclidean distance. These results clarify the relationship between Euclidean distance and patient status, and the lack of motivation for the measurement of density due to its correlation with Euclidean distance. This is a sharper finding, showing that patients robustly have a more reduced vowel space than controls, a difference that was previously only due to 13 patients in Hogoboom et al.'s work.

4.2.2 Correlations with Clinical Research Rating Scales

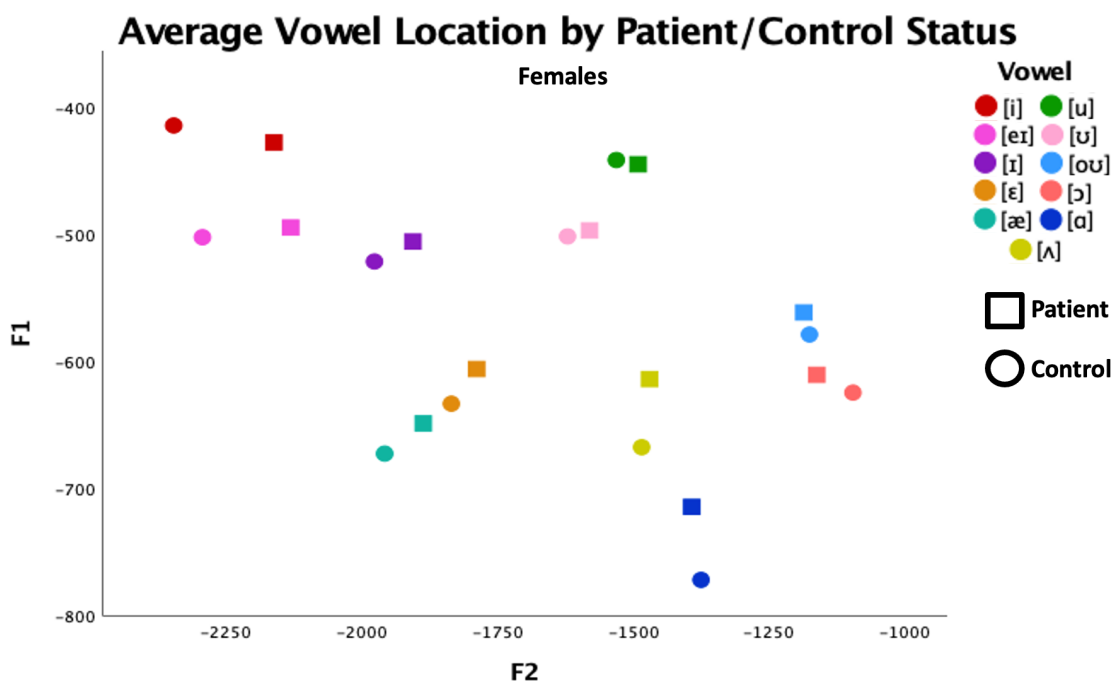
The lack of correlation between aprosody ratings and Euclidean distance is consistent with what Hogoboom et al. found. This illustrates that these underlying phonetic differences in vowel space utilization by patients and controls are not currently being accounted for within the most probable items to encapsulate such linguistic attributes in three of schizophrenia's negative symptom scales.

5. Further Analysis of Vowel Quality

The analysis tool employed by Hogoboom et al. was unable to distinguish between different types of vowels, whereas FAVE relied on separating vowels based on quality and stress. Because of this, we had much richer data to work with than was previously available that allowed us to gain insights into differences in specific types of vowels between patients and controls.

Figures 6 and 7 illustrate the patient and control average stressed vowel measurements, in Hertz, based on gender.

Figure 6. Average stressed vowel measurements for female patients and controls.



As shown above in Figure 6, female controls, represented by the circles, were usually further outside the patient squares, especially on more edge vowels. We see this same trend in the male vowel space, as illustrated by Figure 7, but it is less clear due in males to their smaller vowel space.

Figure 7. Average stressed vowel measurements for male patients and controls.

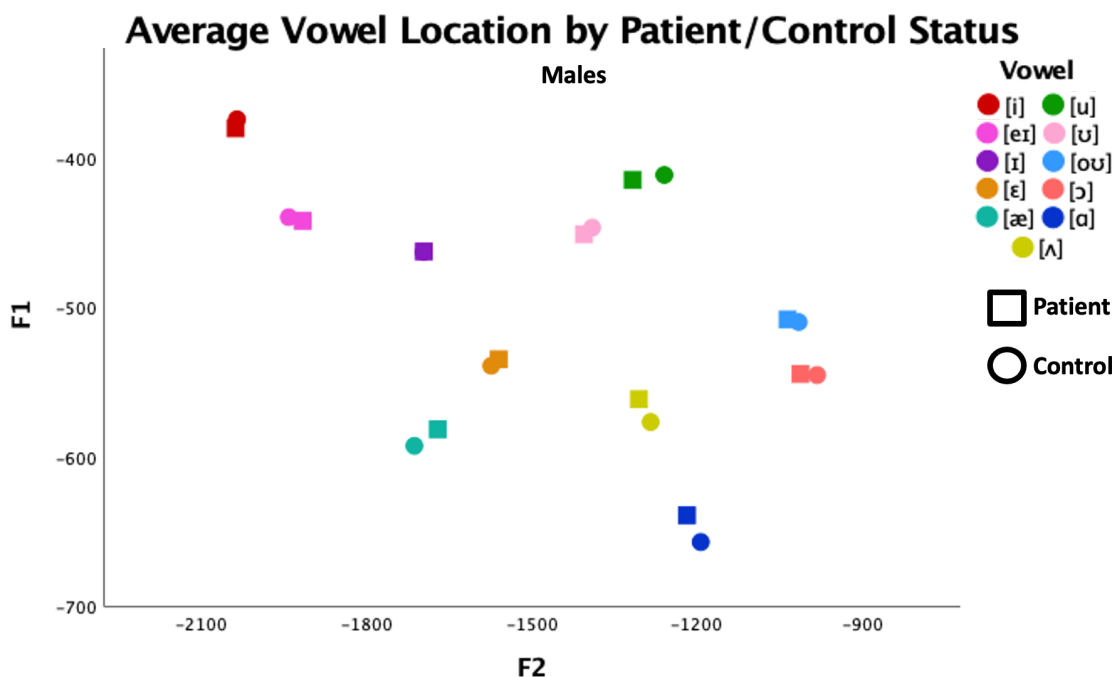


Table 6 shows the frequency of each vowel's occurrence based on patient status; this chart only includes those that met the criteria for analysis described in Section 3.6, separating vowels based on stress where applicable.

Table 6. Frequency of each vowel based on stress and patient status.

Code	IPA Equivalent	Patient Count	Control Count	Total Count
AA1	[ɑ]	1594	1724	3318
AE1	[æ]	3786	4224	8010
AH0	[ə]/[ʌ]	9118	10614	19732
AH1		4068	4062	8130
AO1	[ɔ]	1926	2057	3983
EH1	[ɛ]	2599	2836	5435
EY1	[eɪ]	2021	2170	4191
IH0	[ɪ]	1413	1617	3030
IH1	[ɪ]	3969	4393	8362
IY0	[i]	1313	1443	2756
IY1		3208	3445	6653
OW0	[oʊ]	371	415	786
OW1		2362	2611	4973
UH1	[ʊ]	1042	1139	2181
UW1	[u]	1471	1446	2917
Totals		40331	44270	84601

We ran a generalized linear model with the dependent variable Euclidean distance to determine the effect of the specific vowel on ED. We found that depending on the sound, ED was affected more; for example, patients had a more reduced Euclidean distance for [a] sounds than for [ʌ] sounds. We found that there was not evidence of a significant interaction between the type of vowel produced and the individual's status as a patient or control, as shown in Table 7.

Table 7. Generalized linear regression model effects on Euclidean distance by patient status and vowel.

Euclidean Distance	Wald χ^2	<i>df</i>	Significance
(Intercept)	35646.118	1	<0.001
Sound	4328.641	14	0.001
Patient Status	80.614	1	<0.001
Sound * Patient Status	12.739	14	0.547

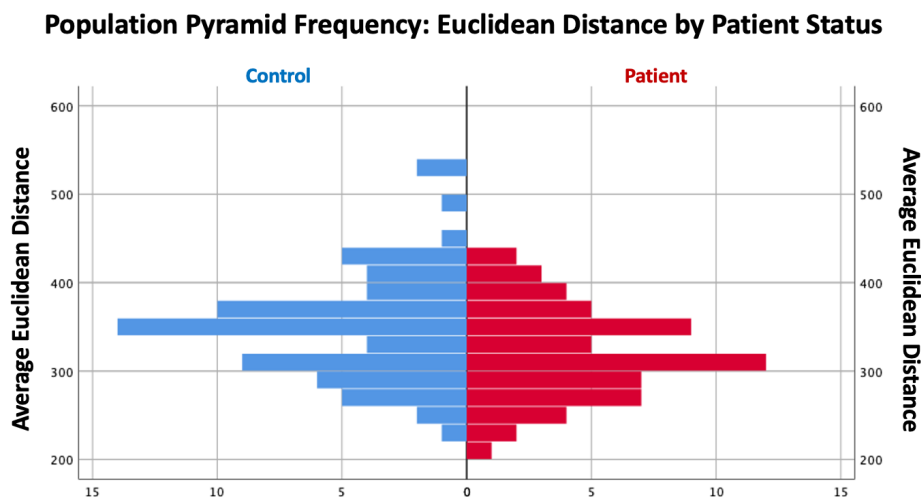
To determine if any demographic factors were impacting our findings, additional Spearman's rho two-tailed bivariate correlations were run on Euclidean distance and demographic variables. There was not a significant correlation between Euclidean distance and these factors, as illustrated in Table 8.

Table 8. Spearman's rho correlation of Euclidean distance and demographic factors.

Euclidean Distance	Age	Ethnicity	Race
Correlation Coefficient	0.110	-0.086	-0.077
Significance (2-tailed)	0.214	0.330	0.387
N	129	129	129

Further analysis compared the average Euclidean distance by patient status, as shown in Figure 8. Upon visual inspection of this data, while there is substantial overlap between the two groups, it appears that if an individual had a Euclidean distance lower than 270 Hz, they are likely to be a patient with schizophrenia.

Figure 8. Frequency of mean Euclidean distance (Hz) based on patient status.



To provide insight to each individual's vowel space variability, Table 9 shows some of the various vowel spaces of participants. Putting gender aside, this highlights the largest and smallest vowel spaces present in the dataset.

Table 9. Largest and smallest vowel spaces based on average Euclidean distance.

	Patient	Control
Smallest Euclidean Distance	<p style="text-align: right;">ED: 207.77 Hz</p>	<p style="text-align: right;">ED: 235.18 Hz</p>
Largest Euclidean Distance	<p style="text-align: right;">ED: 436.36 Hz</p>	<p style="text-align: right;">ED: 529.65 Hz</p>

The lowest overall ED of 207.77 Hz was seen in a patient, whereas the highest overall ED of 529.65 Hz was found in a control. The lowest control ED was 235.18 Hz and the highest patient ED was 436.36 Hz. These examples illustrate a sampling of the wide variety of vowel spaces that exist; we wouldn't expect these differences in vowel spaces to be audibly differentiated between by individual listeners.

6. Discussion

The current analysis finds a strong linguistic measure of Euclidean distance being smaller for patients than controls. This is not currently being accounted for through diagnostic tools, as seen by the lack of correlation with the SANS, CAINS, and PANSS ratings. This measurable, quantifiable difference is one that could be capitalized upon within the diagnostic process to add more reliability and validity to the assessments used in diagnosing schizophrenia. Aspects of speech are able to be quantified, so being able to apply this to the symptomatology of schizophrenia could be useful in better assessing the effects of the disorder on an individual's speech.

It is important to note that not every person with schizophrenia will present with negative symptoms. As noted in Section 1.1, there is a wide array of symptoms that can lead to a schizophrenia diagnosis. This particular research focused on the vowel space correlations with negative symptoms and patient status. The lack of correlation between these negative symptom scale ratings and Euclidean distance is unsurprising, as the scores a patient receives on these tests are subjective to individual clinical listeners, despite supposedly being standardized measures. After an interview with the patient, the clinician fills out many forms rating a person on each of the different items within the scales; to expect the clinician to make precise notes of how an individual did with each item on the scales is not the most feasible approach.

There is a history of perceived honesty biases by clinicians that result in some people being misdiagnosed with schizophrenia. A clinician's distrust of the person they are interviewing during a schizophrenia diagnosis interview has been shown to be a mediator of racial disparities within schizophrenia diagnoses (Eack et al. 2012). A schizophrenia diagnosis has been shown to have pervasive impacts, even when someone with schizophrenia goes to the emergency room. In one study, researchers found that when patients with schizophrenia went to the ER with complaints of pain, clinicians often dismissed the pain due to the person's schizophrenia diagnosis (Shefer et al. 2014). This is one of many issues within the healthcare field regarding how clinical judgments impact patients; therefore, it is crucial to assess how patients are being diagnosed in the first place to avoid having individuals face these issues when the diagnosis is unwarranted.

The way people with schizophrenia are treated by the medical system has extensive room for improvement, particularly by adjusting the hierarchical structure of psychological disorders. For starters, a variety of psychiatric disorders, including depression and OCD, co-occur with schizophrenia, but the overall psychiatric diagnostic process has a hierarchical structure that

blocks these co-occurring disorders from being diagnosed and effectively treated (Bermanzohn et al. 2000). The hierarchical structure operates in a way that ranks certain disorders above others; those higher up are seen as encompassing the symptoms of the lower disorders, thus resulting in a schizophrenia diagnosis, which is more highly ranked, being seen as involving all of the different symptoms of diagnoses below it in the hierarchy (Surtees & Kendell 1979). Despite the fact that some people meet the criteria for certain lower-ranked disorders, patients do not receive these diagnoses that could be effectively treated; while it is incredibly difficult to disambiguate symptoms from their cause, the flexibility within the *DSM* has not been effectively altered to allow for the formal diagnosis of these disorders in tandem with schizophrenia as often as is likely necessary (Bermanzohn et al. 2000).

There is also a history of racism and distrust within medicine that extends into mental health services. The Black population is often subjected to microaggressions and racial prejudice by a white-dominated provider population. Racial factors have been shown to be influential in the diagnostic stage of schizophrenia in particular. For example, one study looked at how clinicians of different races made decisions regarding the diagnosis of patients; they found that non-African American clinicians frequently associated negative symptoms with a schizophrenia diagnosis, whereas the African American clinicians did not (Trierweiler et al. 2006). Additionally, African American clinicians would diagnose schizophrenia at lower rates when they considered substance abuse issues, revealing a stark difference in how an individual's symptoms are analyzed and diagnosed, simply on the basis of individual clinical listener (Trierweiler et al. 2006). African American clinicians more frequently actively considered situational circumstances and information, and while that did not correlate to any changes in diagnosis, researchers found that when non-African American clinicians took that situational information into consideration, they were more likely to diagnose a patient with a mood disorder rather than schizophrenia (Trierweiler et al. 2005). Furthermore, a separate study looked at the psychiatric interview process and found that white patients were more likely to be diagnosed with a mood disorder than Black patients; the semi-structured interview process within the *DSM-III* was found to diagnose a higher percentage of Black patients with schizophrenia than white patients (Neighbors et al. 1999). This disparity was reflected in additional work that found that white individuals were more likely to be diagnosed with bipolar disorder than Black individuals, who were, again, more likely to receive the schizophrenia diagnosis (Neighbors et al. 2003). Moreover, researchers found that, after controlling for indicators of SES at birth, African American patients were twice as likely to be diagnosed with schizophrenia as their white counterparts (Bresnahan et al. 2007). It is important to note that despite the lack of correlation with ED and race in the current study, we cannot rule out these biases from having affected the data source.

As previously mentioned, schizophrenia has a wide variety of symptoms that are challenging to measure and occur at different rates by individual; still, racial biases influence the ambiguity of the diagnostic process. One study noted that a positive symptom, like hallucinations, that was present in different rates between a group of African American patients

and a group of white patients did not result in different rates of a schizophrenia diagnosis; instead, clinicians attributed negative symptoms, which occurred equally between groups, with a schizophrenia diagnosis at higher rates within the African American patient group than the white patient group (Trierweiler et al. 2000). While efforts have been made to increase flexibility and limit the impact of individual clinician bias on the diagnosis of patients with schizophrenia, studies have shown that these structured assessments fail to fully mitigate these influences (Olbert et al. 2018). The *DSM-V* made some changes in small details of how negative symptoms are described, but did not make any adjustments to how they are measured (Reddy et al. 2014).

Moreover, mental health providers have been finding many instances of inaccurate schizophrenia diagnoses (Tzur Bitan et al. 2018). This has been shown to be pervasive within psychological diagnoses, including schizophrenia. One study gathered data from the state of Indiana and found that African American patients were more often diagnosed with schizophrenia than white patients; after they controlled for other demographic variables, they found that race was the strongest predictor of being admitted to the state psychiatric hospital under a schizophrenia diagnosis (Barnes 2008). Additionally, the *DSM-V* has been shown to perpetuate these racial differences through the semi-structured interviews, thus prompting an increased reliance on clinical judgments (Neighbors et al. 2003). This overdiagnosis of schizophrenia within the Black community has implications that extend beyond a simple diagnosis; therefore, it is crucial that some more quantifiable dimensions of schizophrenia are pinned down in order to limit clinician bias.

Analyzing the vowel space of individuals in relation to psychological diagnoses has been done, like with depression and PTSD. While these studies did not utilize Euclidean distance, they found that there was a reduced vowel space for individuals with depression and PTSD (Scherer et al. 2016). The means by which they came to this conclusion were not as linguistically-informed as current tools allow; rather, they utilized three vowels as reference points (/i/, /a/, and /u/) and took ratios to determine that a vowel space was reduced. Another study found that a reduced vowel space correlated with signs of psychological distress (Scherer et al. 2015). The results from this present study regarding Euclidean distance have the potential to be used as an approach to analyze speech for more disorders than only schizophrenia. While it may not clarify differences between depression and schizophrenia, it could at least be a factor that limits the number of unaffected individuals being misdiagnosed with schizophrenia.

Certain aspects of this study were out of our control, such as the selected reading and the interruptions, including doors slamming, phones ringing, and additional people whispering, that were present in some of the recordings. The reading was an excerpt from the children's fiction book *Heidi* by Johanna Spyri, written in British English (1998 [1880]). It is harder to read a dialect that one does not speak, so phrases like, "and it was well it was so," and, "Heidi [...] run from one window to the other," caused many people to stumble a bit because they were unnatural utterances for speakers of American English. While it was more systematic to analyze speech from Task 4 due to its nature as a reading-based study with predictable transcriptions, speech produced from reading aloud is not the same as natural speech. Ideally, one would

transcribe spontaneous speech, but the amount of data produced would vary by participant and the distribution of sounds would be harder to standardize between groups.

While this study found that density was correlated with Euclidean distance, additional research is required to assess if density, as proposed by Hogoboom et al. (submitted), is a useful tool for analyzing the vowel space and captures a measurable difference that simply was not relevant to this dataset. Additionally, further work could be done to assess if there is an underlying significant interaction between gender and patient status that was limited by the sample size of this study.

While it is currently out of reach, this research could lend itself toward creating more automated diagnostic tools, providing clinicians with more consistent ways to measure potential aspects of schizophrenia than basing it on their past experiences with other patients and any influence of their personal racial biases. Future studies could assess newer fully-automated forced aligners, like DARLA, in comparison to FAVE and individual transcribers to see if these measurable differences in vowel formants are maintained by the fully-automated forced aligners. If so, this could make it easier to implement speech analysis at the clinical level, as there would not be the need for an individual to transcribe each word. Additional work still needs to be done to increase the accuracy of transcriptions and vowel measurements of DARLA, but these advancements provide promising outlooks for the increased utilization of linguistically-informed speech analysis within medicine. By looking into the characteristics of speech, we could find additional quantifiable means by which to assess patients, ultimately providing clinicians with more robust measures to assess psychological disorders that are less likely to be affected by individual clinician judgments and biases.

Finally, a useful tool that could be made for phonetic transcriptions in general could combine the Praat spectrograms with the ELAN transcription functions. This would allow for researchers to look more closely at the spectrograms in order to discern any sounds that were hard to determine from the recording, as the formants present in spectrograms are useful tools for differentiating between sound qualities. This would save time and increase the accuracy of phonetic transcriptions, as one could ultimately narrowly transcribe more accurately based on what the formants indicate about the different sounds, rather than referring back to Praat each time a sound needs to be clarified.

7. Conclusion

Our work expanded upon previous work of quantifying the vowel space of patients with schizophrenia and controls, clarifying that there is a relationship between vowel space dispersion, on the basis of Euclidean distance, and patient status. While Hogoboom et al. (submitted) found ED to be a weak predictor of patient status and that Euclidean distance and vowel density were, in tandem, significant predictors of patient status, this study found that ED and vowel density were correlated and that ED was a strong predictor of patient status ($p=0.005$). This difference in Euclidean distance is not currently being accounted for in the psychological scales used to assess schizophrenia severity.

Clinician bias and racism within the medical field are pervasive, resulting primarily in members of the Black community being overdiagnosed with schizophrenia. This misdiagnosis carries heavy implications that impact their personal lives and influence their future medical encounters, so it is crucial that less subjective and more quantitative measures, like the use of forced aligners to gather vowel measurements, are utilized to mitigate the influence of these biases.

Appendix A

Item Descriptions of Negative Symptoms of Schizophrenia

SANS 7

Lack of Vocal Inflections (within “Affective Flattening or Blunting”)

“The patient fails to show normal vocal emphasis patterns, is often monotonic.”

SANS 8

Global Rating of Affective Flattening (within “Affective Flattening or Blunting”)

“This rating should focus on overall severity of symptoms, especially unresponsiveness, inappropriateness, and an overall decrease in emotional intensity.”

CAINS 11

Vocal Expression (within “Expression Scale”)

“This item refers to prosodic features of the voice. This item reflects changes in tone during the course of speech. Speech rate, amount, or content of speech is not assessed.”

PANSS N1

Blunted Affect (within “Negative Scale”)

“Diminished emotional responsiveness as characterized by a reduction in facial expression, modulation of feelings, and communicative gestures. Basis for rating: observation of physical manifestation of affective tone and emotional responsiveness during the course of interview.”

Appendix B

Heidi Excerpt from Chapter IV: “The Visit to Grandmother” (Spyri 1998 [1880]).

The thing which attracted her most, however, was the waving and roaring of the three old fir trees on these windy days. She would run away repeatedly from whatever she might be doing, to listen to them, for nothing seemed so strange and wonderful to her as the deep mysterious sound in the tops of the trees. She would stand underneath them and look up, unable to tear herself away, looking and listening while they bowed and swayed and roared as the mighty wind rushed through them. There was no longer now the warm bright sun that had shone all through the summer, so Heidi went to the cupboard and got out her shoes and stockings and dress, for it was growing colder every day, and when Heidi stood under the fir trees the wind blew through her as if she was a thin little leaf, but still she felt she could not stay indoors when she heard the branches waving outside.

Then it grew very cold, and Peter would come up early in the morning blowing on his fingers to keep them warm. But he soon left off coming, for one night there was a heavy fall of snow and the next morning the whole mountain was covered with it, and not a single little green leaf was to be seen anywhere upon it. There was no Peter that day, and Heidi stood at the little window looking out in wonderment, for the snow was beginning again, and the thick flakes kept falling till the snow was up to the window, and still they continued to fall, and the snow grew higher, so that at last the window could not be opened, and she and her grandfather were shut up fast within the hut. Heidi thought this was great fun and ran from one window to the other to see what would happen next, and whether the snow was going to cover up the whole hut, so that they would have to light a lamp although it was broad daylight. But things did not get as bad as that, and the next day, the snow having ceased, the grandfather went out and shovelled away the snow round the house, and threw it into such great heaps that they looked like mountains standing at intervals on either side the hut. And now the windows and door could be opened, and it was well it was so, for as Heidi and her grandfather were sitting one afternoon on their three-legged stools before the fire there came a great thump at the door followed by several others, and then the door opened. It was Peter, who had made all that noise knocking the snow off his shoes; he was still white all over with it, for he had had to fight his way through deep snowdrifts, and large lumps of snow that had frozen upon him still clung to his clothes. He had been determined, however, not to be beaten and to climb up to the hut, for it was a week now since he had seen Heidi.

"Good-evening," he said as he came in; then he went and placed himself as near the fire as he could without saying another word, but his whole face was beaming with pleasure at finding himself there. Heidi looked on in astonishment, for Peter was beginning to thaw all over with the warmth, so that he had the appearance of a trickling waterfall.

References

- Bailey, George. 2016. Automatic detection of sociolinguistic variation using forced alignment. *University of Pennsylvania Working Papers in Linguistics* 22(2). 10-20. <https://repository.upenn.edu/pwpl/vol22/iss2/3>
- Barnes, Arnold. 2008. Race and hospital diagnoses of schizophrenia and mood disorders. *Social Work* 53(1). 77-83. doi.org/10.1093/sw/53.1.77
- Bermanzohn, Paul C., Linda Porto, Phyllis B. Arlow, Simcha Pollack, Roslyn Stronger, & Samuel G. Siris. 2000. Hierarchical diagnoses in chronic schizophrenia: A clinical study of co-occurring syndromes. *Schizophrenia Bulletin* 26(3). 517-525. doi.org/10.1093/oxfordjournals.schbul.a033472
- Boersma, Paul & David Weenink. 2019. Praat: Doing phonetics by computer. [Computer Program]. Version 6.0.48. Retrieved from www.praat.org.
- Bresnahan, Michaeline, Melissa D. Begg, Alan Brown, Catherine Schaefer, Nancy Sohler, Beverly Insel, Leah Vella, & Ezra Susser. 2007. Race and risk of schizophrenia in a US birth cohort: Another example of health disparity? *International Journal of Epidemiology* 36(4). 751-758. doi.org/10.1093/ije/dym041

- Clopper, Cynthia G. & Janet B. Pierrehumbert. 2008. Effects of semantic predictability and regional dialect of vowel space reduction. *The Journal of the Acoustical Society of America* 124(3). 1682-1688. doi.org/10.1121/1.2953322
- Clopper, Cynthia G., Jane F. Mitsch, & Terrin N. Tamati. 2016. Effects of phonetic reduction and regional dialect on vowel production. *Journal of Phonetics* 60. 38-59. doi.org/10.1016/j.wocn.2016.11.002
- Compton, Michael T., Anya Lunden, Sean D. Cleary, Luca Pauselli, Yazeed Alolayan, Brooke Halpern, Beth Broussard, Anthony Crisafio, Leslie Capulong, Pierfrancesco Maria Balducci, Francesco Bernardini, & Michael A. Covington. 2018. The aprosody of schizophrenia: Computationally derived acoustic underpinnings of monotone speech. *Schizophrenia Research* 197. 392-399. doi.org/10.1016/j.schres.2018.01.007
- Covington, Michael A., Anya Lunden, Sarah L. Cristofaro, Claire Ramsay Wan, C. Thomas Bailey, Beth Broussard, Robert Fogarty, Stephanie Johnson, Shayl Zhang, & Michael T. Compton. 2012. Phonetic measures of reduced tongue movement correlate with negative symptom severity in hospitalized patients with first-episode schizophrenia-spectrum disorders. *Schizophrenia Research* 142(1-3). 93-95. doi.org/10.1016/j.schres.2012.10.005
- Craig, Holly K., Connie A. Thompson, Julie A. Washington, & Stephanie L. Potter. 2003. Phonological features of child African American English. *Journal of Speech, Language, and Hearing Research* 46. 623-635. doi.org/10.1044/1092-4388(2003/049)
- Eack, Shaun M., Amber L. Bahorik, Christina E. Newhill, Harold W. Neighbors, & Larry E. Davis. 2012. Interviewer-perceived honesty as a mediator of racial disparities in the diagnosis of schizophrenia. *Psychiatric Services* 63(9). 875-880. doi.org/10.1176/appi.ps.201100388
- ELAN (Version 6.0) [Computer software]. 2020. Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. Retrieved from archive.mpi.nl/tla/elan
- Evanini, Keelan, Stephen Isard, & Mark Liberman. 2009. Automatic formant extraction for sociolinguistic analysis of large corpora. *Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, September 6-10, Brighton, UK, 1655-1658.
- Fox, Robert Allen & Ewa Jacewicz. 2009. Cross-dialectal variation in formant dynamics of American English vowels. *The Journal of the Acoustical Society of America* 126. 2603-2618. doi.org/10.1121/1.3212921
- Hinzen, Wolfram & Joana Rosselló. 2015. The linguistics of schizophrenia: Thought disturbance as language pathology across positive symptoms. *Frontiers in Psychology* 6. 1-17. doi.org/10.3389/fpsyg.2015.00971
- Hogoboom, Anya, Megan Rouch, Diana Worthen, Luca Pauselli, & Michael T. Compton. Submitted. Vowel space reduction as an indicator of negative symptoms in a subset of individuals with schizophrenia.
- Kay, Stanley R., Abraham Fiszbein, & Lewis A. Opler. 1987. The Positive and Negative Syndrome Scale (PANSS) for schizophrenia. *Schizophrenia Bulletin* 13(2). 261-276. doi.org/10.1093/schbul/13.2.261
- Kumari, Suneeta, Mansoor Malik, Christina Florival, Partam Manalai, & Snezana Sonje. 2017. An assessment of five (PANSS, SAPS, SANS, NSA-16, CGI-SCH) commonly used symptoms rating scales in schizophrenia and comparison to newer scales (CAINS, BNSS). *Journal of Addiction Research and Therapy* 8(3). doi.org/10.4172/2155-6105.1000324

- Lennes, Mietta. 2003. Praat script. (Modified by Dan McCloy, December 2011).
- Lippi-Green, Rosina. 2011. *English with an Accent: Language, ideology, and discrimination in the United States*. Taylor & Francis Group.
- Low, Daniel M., Kate H. Bentley, & Satrajit S. Ghosh. 2020. Automated assessment of psychiatric disorders using speech: a systematic review. *Laryngoscope Investigative Otolaryngology* 5(1). 96-116. doi.org/10.1002/lio2.354
- MacKenzie, Laurel & Danielle Turton. 2020. Assessing the accuracy of existing forced alignment software on varieties of British English. *Linguistics Vanguard* 6(s1). doi.org/10.1515/lingvan-2018-0061
- Meer, Philipp, Thorsten Brato, & José Alejandro Matute Flores. 2021. Extending automatic vowel formant extraction to New Englishes. *English World-Wide* 42(1). 54-84. doi.org/10.1075/eww.00060.mee
- Meer, Philipp. 2020. Automatic alignment for New Englishes: Applying state-of-the-art aligners to Trinidadian English. *Journal of the Acoustical Society of America* 147(4). 2283-2294. doi.org/10.1121/10.0001069
- Mertens, Piet. 2014. Polytonia: A system for the automatic transcription of tonal aspects in speech corpora. *Journal of Speech Sciences* 4(2). 17-57.
- Murphy, Brendan P., Yung-Chul Chung, Tae-Won Park, & Patrick D. McGorry. 2006. Pharmacological treatment of primary and negative symptoms in schizophrenia: A systematic review. *Schizophrenia Research* 88(1-3). 5-25. doi.org/10.1016/j.schres.2006.07.002
- Neighbors, Harold W., Steven J. Trierweiler, Cheryl Munday, Estina E. Thompson, James S. Jackson, Victoria J. Binion, & John Gomez. 1999. Psychiatric diagnosis of African Americans: Diagnostic divergence in clinician-structured and semistructured interviewing conditions. *Journal of the National Medical Association* 91(11). 601-612.
- Neighbors, Harold W., Steven J. Trierweiler, Briggett C. Ford, & Jordana R. Muroff. 2003. Racial differences in DSM diagnosis using a semi-structured instrument: The importance of clinical judgment in the diagnosis of African Americans. *Journal of Health and Social Behavior* 44(3). 237-256. doi.org/10.2307/1519777
- Nycz, Jennifer & Lauren Hall-Lew. 2013. Best practices in measuring vowel merger. *The Journal of the Acoustical Society of America* 20(1). doi.org/10.1121/1.4894063
- Olbert, Charles M., Arundati Nagendra, & Benjamin Buck. 2018. Meta-analysis of Black vs white racial disparity in schizophrenia diagnosis in the United States: Do structured assessments attenuate racial disparities? *Journal of Abnormal Psychology* 127(1). 104-115. doi.org/10.1037/abn0000309
- Povey, Daniel, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, & Karel Vesely. 2011. The Kaldi speech recognition toolkit. *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society
- Reddy, L. Felice, William P. Horan, & Michael F. Green. 2014. Revisions and refinements of the diagnosis of schizophrenia in DSM-5. *Clinical Psychology: Science and Practice* 21(3). 236-244. doi.org/10.1111/cpsp.12071
- Reddy, Sravana & James N. Stanford. 2015. Toward completely automated vowel extraction: Introducing DARLA. *Linguistics Vanguard* 1(1). 15-28. doi.org/10.1515/lingvan-2015-0002

- Rosenfelder, Ingrid, Josef Fruehwald, Keelan Evanini, Scott Seyfarth, Kyle Gorman, Hilary Prichard, Jiahong Yuan. 2014. FAVE (Forced Alignment and Vowel Extraction) Program Suite v1.2.2 10.5281/zenodo.22281
- Scherer, Stefan, Gale M. Lucas, Jonathan Gratch, Albert "Skip" Rizzo, & Louis-Philippe Morency. 2016. 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. doi.org/10.1109/TAFFC.2015.2440264
- Scherer, Stefan, Louis-Philippe Morency, Jonathan Gratch, & John Pestian. 2015. Reduced vowel space is a robust indicator of psychological distress: A cross-corpus analysis. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 4789-4793. doi.org/10.1109/ICASSP.2015.7178880
- Shefer, Guy, Claire Henderson, Louise M. Howard, Joanna Murray, & Graham Thornicroft. 2014. Diagnostic overshadowing and other challenges involved in the diagnostic process of patients with mental illness who present in emergency departments with physical symptoms - a qualitative study. *PLoS ONE* 9(11). 1-8. doi.org/10.1371/journal.pone.0111682
- Spyri, Johanna. (1998 [1880]). Heidi. Urbana, Illinois: Project Gutenberg. Retrieved March 12, 2021, from gutenberg.org/cache/epub/1448/pg1448.html
- Strauss, Gregory P. & James M. Gold. 2016. A psychometric comparison of the Clinical Assessment Interview for Negative Symptoms and the Brief Negative Symptom Scale. *Schizophrenia Bulletin* 42(6). 1384-1394. doi.org/10.1093/schbul/sbw046
- Surtees, P. G. & R. E. Kendell. 1979. The hierarchy model of psychiatric symptomatology: An investigation based on present state examination ratings. *The British Journal of Psychiatry: The Journal of Mental Science* 135. 438-443. doi.org/10.1192/bjp.135.5.438
- Tahir, Yasir, Zixu Yang, Debsubhra Chakraborty, Nadia Thalmann, Daniel Thalmann, Yogeswary Maniam, Nur Amirah binte Abdul Rashid, Bhing-Leet Tan, Jimmy Lee Chee Keong, & Justin Dauwels. 2019. Non-verbal speech cues as objective measures for negative symptoms in patients with schizophrenia. *PLoS ONE* 14(4). 1-17. doi.org/10.1371/journal.pone.0214314
- Tandon, Rajiv, Wolfgang Gaebel, Deanna M. Barch, Juan Bustillo, Raquel E. Gur, Stephan Heckers, Dolores Malaspina, Michael J. Owen, Susan Schultz, Ming Tsuang, Jim Van Os, & William Carpenter. 2013. Definition and description of schizophrenia in the DSM-5. *Schizophrenia Research* 150(1). 3-10. doi.org/10.1016/j.schres.2013.05.028
- Trierweiler, Steven J., Harold W. Neighbors, Cheryl Munday, Estina E. Thompson, Victoria J. Binion, & John P. Gomez. 2000. Clinician attributions associated with the diagnosis of schizophrenia in African American and non-African American patients. *Journal of Consulting and Clinical Psychology* 68(1). 171-175. doi.org/10.1037/0022-006X.68.1.171
- Trierweiler, Steven J., Harold W. Neighbors, Cheryl Munday, Estina E. Thompson, James S. Jackson, & Victoria J. Binion. 2006. Differences in patterns of symptom attribution in diagnosing schizophrenia between African American and non-African American clinicians. *American Journal of Orthopsychiatry* 76(2). 154-160. doi.org/10.1037/0002-9432.76.2.154
- Trierweiler, Steven J., Jordana R. Muroff, James S. Jackson, Harold W. Neighbors, & Cheryl Munday. 2005. Clinician race, situational attributions, and diagnoses of mood versus

- schizophrenia disorders. *Cultural Diversity and Ethnic Minority Psychology* 11(4). 351-364. doi.org/10.1037/1099-9809.11.4.351
- Tzur Bitan, Dana, Ariella Grossman Giron, Gady Alon, Shlomo Mendlovic, Yuval Bloch, & Aviv Segev. 2018. Attitudes of mental health clinicians toward perceived inaccuracy of a schizophrenia diagnosis in routine clinical practice. *BMC Psychiatry* 18. doi.org/10.1186/s12888-018-1897-2
- Yang, Byunggon & D. H. Whalen. 2015. Perception and production of English vowels by American males and females. *Australian Journal of Linguistics* 35(2). 121-141. doi.org/10.1080/07268602.2015.1004998
- Yuan, Jiahong & Mark Liberman. 2008. Speaker identification on the SCOTUS corpus. *The Journal of the Acoustical Society of America* 123(5). doi.org/10.1121/1.2935783