Vowel Acoustics as Predictors of Speech Intelligibility in Dysarthria

Austin Thompson,[a] [iD] Micah E. Hirsch,[a] [iD] Kaitlin L. Lansford,[a] [iD] and Yunjung Kim[a] [iD]

[a] School of Communication Science and Disorders, Florida State University, Tallahassee

ABSTRACT

Purpose: This study sought to determine if alternative vowel space area (VSA) measures (i.e., novel trajectory-based measures: vowel space hull area and vowel space density) predicted speech intelligibility to the same extent as two traditional vowel measures (i.e., token-based measures: VSA and corner dispersion) in speakers with dysarthria. Additionally, this study examined if the strength of the relationship between acoustic vowel measures and intelligibility differed based on how intelligibility was measured (i.e., orthographic transcriptions [OTs] and visual analog scale [VAS] ratings).
Method: The Grandfather Passage was read aloud by 40 speakers with dysarthria of varying etiologies, including Parkinson's disease ($n = 10$), amyotrophic lateral sclerosis ($n = 10$), Huntington's disease ($n = 10$), and cerebellar ataxia ($n = 10$). Token- and trajectory-based acoustic vowel measures were calculated from the passage. Naïve listeners ($N = 140$) were recruited via crowdsourcing to provide OTs and VAS intelligibility ratings. Hierarchical linear regression models were created to model OTs and VAS intelligibility ratings using the acoustic vowel measures as predictors.
Results: Traditional VSA was the sole significant predictor of speech intelligibility for both the OTs ($R^2 = .259$) and VAS ($R^2 = .236$) models. In contrast, the trajectory-based measures were not significant predictors of intelligibility. Additionally, the OTs and VAS intelligibility ratings conveyed similar information.
Conclusions: The findings suggest that traditional token-based vowel measures better predict intelligibility than trajectory-based measures. Additionally, the findings suggest that VAS methods are comparable to OT methods for estimating speech intelligibility for research purposes.

Acoustic vowel space area (VSA) is a widely studied measure of articulatory working space used to investigate vowel degradation in dysarthria. Most commonly, this measurement is calculated as the planar area of the first and second formant frequency (F1 and F2, respectively) information extracted from the temporal midpoint of four corner vowels in American English (i.e., /i/, /u/, /ɑ/, and /æ/; Kent & Y-J. Kim, 2003). An expanded acoustic VSA indicates greater acoustic distinctiveness between corner vowels' articulatory positions, which infers larger articulatory movement.

In contrast, a reduced VSA indicates less acoustic distinctiveness, likely due to smaller articulatory movement (Weismer et al., 2012). As a measure, VSA has been evidenced to have two primary applications for dysarthria: (a) detecting the presence of dysarthria and (b) predicting perceptual outcomes, namely, speech intelligibility. These applications are discussed in greater detail below.

Speakers with dysarthria are almost universally found to have reduced VSA compared to neurologically healthy speakers, suggesting a decreased articulatory working space and less perceptually distinct vowels. Relative to neurologically healthy speakers, reduced VSA has been reported in speakers with Parkinson's disease (PD; Lam & Tjaden, 2016; McRae et al., 2002; Tjaden et al., 2013; Tjaden & Wilding, 2004), amyotrophic lateral sclerosis (ALS; Lansford & Liss, 2014b; Turner et al., 1995), ataxia

(Lansford & Liss, 2014b), Huntington's disease (HD; Lansford & Liss, 2014b; Rusz et al., 2014), and cerebral palsy (Higgins & Hodge, 2002). Importantly, VSA is not sensitive to differences between dysarthria subtypes (Y-J. Kim et al., 2011; Lansford & Liss, 2014b). Instead, reduced VSA is likely a global feature of dysarthria and, as a result, may be a reliable acoustic biomarker of the presence of dysarthria in clinical populations.

Intelligibility is a multidimensional construct sensitive to breakdowns in all subsystems of speech production, including articulatory imprecision. Given the presumed negative consequences of vowel reduction on producing perceptually distinct vowels, it follows that VSA may also serve as an index for intelligibility deficits in dysarthria. Indeed, previous research supports this assumption, with reports of VSA being positively related to speech intelligibility (Lansford & Liss, 2014a; Lee et al., 2016; Turner et al., 1995; Weismer et al., 2001). However, the strength of the relationship between VSA and intelligibility varies greatly across experiments (see Table 1). For example, Lee et al. (2016) reported a strong and positive relationship between VSA and speech intelligibility ($R^2 = .79$) in speakers with dysarthria due to ALS ($n = 11$) and neurologically healthy control speakers ($n = 11$). In contrast, Turner et al. (1995) reported a weaker, albeit moderate and positive, relationship between VSA and intelligibility ($R^2 = .45$) in speakers with ALS ($n = 9$) and neurotypical speakers ($n = 9$). The different magnitudes of the observed relationship between VSA and intelligibility between the two studies may be accounted for by differing methods for measuring intelligibility. Specifically, Lee et al. measured intelligibility using orthographic transcriptions (OTs) obtained from the Sentence Intelligibility Test, whereas

Turner et al. measured intelligibility using a direct magnitude estimation (DME) approach. Together, these findings suggest that the strength of the relationship between VSA and intelligibility may differ based on how intelligibility is measured.

Other investigations of VSA and intelligibility have reported much weaker relationships. For example, Tjaden and Wilding (2004) is an often-cited example of poor relations between VSA and speech intelligibility, with R^2 values ranging from .06 to .08. This study examined the relationships between intelligibility and VSA, first-moment difference measures, and F2 slope for males and females with PD ($n = 12$) and multiple sclerosis ($n = 15$). However, the relatively small R^2 values may partly be caused by the statistical approach used in their study. Specifically, they examined these measures across three speaking conditions (i.e., habitual, loud, and slow). Therefore, their data were nested by speaker (i.e., each speaker had three productions of the target stimuli, one for each condition). However, the authors used linear regression to examine the relationship between these acoustic measures and intelligibility, which does not account for the nested data structure. Instead, a linear mixed-effects analysis of the data would have allowed the authors to account for speaker variance, which would have likely yielded higher R^2 values.

In another study, Neel (2008) found a weak relationship between VSA and intelligibility ($R^2_{\text{Males}} = .12$; $R^2_{\text{Females}} = .09$) in neurologically healthy speakers ($N = 93$). However, this weak relationship may partly be explained by the high intelligibility of the speakers and, therefore, only represents the upper range of the relationship between VSA and intelligibility. Similar weak relationships have been observed in studies investigating speakers

Table 1. Selected studies investigating the relationship between vowel space area (VSA) and speech intelligibility.

Reference	N	Populations	Intelligibility measure	R^2
VSA				
Y-J. Kim et al. (2011)	107	PD, stroke, TBI, MSA	DME	.28
Neel (2008)	48	HC females	FC	.09
Neel (2008)	45	HC males	FC	.12
Lansford & Liss (2014a)	45	PD, ALS, HD, ataxia	OT	.16
Weismer et al. (2001)	39	HC, PD, ALS	DME	.46
McRae et al. (2002)	26	HC, PD	DME	.13
DuHadway & Hustad (2012)	24	HC, CP	OT	.63
Lee et al. (2016)	22	HC, ALS	OT	.79
Turner et al. (1995)	18	HC, ALS	DME	.45
Tjaden & Wilding (2004)	16	PD, MS females	DME	.06–.08
Higgins & Hodge (2002)	12	HC, CP	OT	.41–.53
Tjaden & Wilding (2004)	11	PD, MS males	DME	.12–.21
Triangular VSA				
Liu et al. (2005)	30	HC, PD	OT	.47
H. Kim et al. (2011)	12	HC, CP	OT	.69

Note. PD = Parkinson's disease; TBI = traumatic brain injury; MSA = multiple systems atrophy; DME = direct magnitude estimation; HC = neurologically healthy control speakers; FC = forced-choice vowel identification; ALS = amyotrophic lateral sclerosis; HD = Huntington's disease; OT = orthographic transcription; CP = cerebral palsy; MS = multiple sclerosis.

with mild dysarthria and relatively high intelligibility, such as speakers with PD (McRae et al., 2002). Therefore, when investigating the relationship between VSA and intelligibility, it may be beneficial for investigations to include many speakers representing a wide range of intelligibility.

Only a few studies have met these criteria (Y-J. Kim et al., 2011; Lansford & Liss, 2014a; Weismer et al., 2001). Y-J. Kim et al. (2011), observed a moderate positive correlation between VSA and intelligibility (R^2 = .28) in speakers with dysarthria secondary to PD, stroke, traumatic brain injury, and multiple systems atrophy (N = 107). Similarly, Lansford and Liss (2014a) reported a moderate positive correlation between VSA and intelligibility (R^2 = .16) in speakers with dysarthria due to PD and ALS ataxia, and HD (N = 45). Finally, Weismer et al. (2001) reported a moderate-to-strong positive correlation between VSA and intelligibility (R^2 = .46) in speakers with dysarthria due to PD, ALS, and neurologically healthy control speakers (N = 39). These studies suggest that the relationship between VSA and intelligibility is likely positive and moderate.

However, many researchers have argued that VSA is an imperfect measure. Specifically, researchers have criticized VSA because it uses formant information taken from a snapshot of only four, and sometimes three, corner vowels to make inferences about the size of the articulatory working space (Sandoval et al., 2013; Story & Bunton, 2017). Another limitation of VSA is that dialect heavily impacts the measure, making comparisons across dialects (and even variations within dialects) problematic (Fox & Jacewicz, 2017). This limitation makes studying dialect, an already understudied area within motor speech disorders research, challenging. Finally, others have criticized VSA because its measurement is cumbersome and requires identifying appropriate corner vowel tokens for calculation (Sandoval et al., 2013). These issues have led researchers to propose alternative acoustic measures to quantify vowel distinctiveness, such as corner dispersion (Lansford & Liss, 2014a; Turner et al., 1995), vowel space density (VSD; Story & Bunton, 2017; Whitfield & Mehta, 2019), automatic VSA (Sandoval et al., 2013), vowel space hull (Whitfield et al., 2018; Whitfield & Mehta, 2019), articulatory–acoustic vowel space (Whitfield & Goberman, 2014), formant centralization ratio (Sapir et al., 2010), and vowel articulation index (Skodda et al., 2011).

Corner dispersion, sometimes referred to simply as vowel distance, is an acoustic measure calculated as the Euclidean distance between the corner vowels and the center vowel /ʌ/ (Lansford & Liss, 2014a), or the centroid of the vowel space (Tjaden et al., 2013; Turner et al., 1995). Like the traditional VSA measure, corner dispersion relies on the identification and manual segmentation of specific vowel tokens to obtain formant frequency data (typically from the midpoint of the vowel). However, unlike traditional VSA, corner dispersion allows for the examination of the relative contribution of each corner vowel to the centralization of the VSA, which may provide more nuanced information regarding articulatory deficits in vowel production. Furthermore, like traditional VSA, corner dispersion has been observed to have a moderate and positive relationship to intelligibility, R^2 = .209 (Lansford & Liss, 2014a).

In contrast to token-based measures like traditional VSA and corner dispersion, many of the proposed alternative measures are advantageous because they can be calculated automatically from formant data extracted at the passage or conversational level and do not require manual segmentation of the acoustic signal. For example, VSD is a measure proposed by Story and Bunton (2017) to examine the vowel characteristics produced by speakers over a longer duration, such as passage reading. This measure uses the continuous formant trajectories of voiced speech to calculate the speaker's formant space density. More dense areas of the speaker's formant space reveal "articulatory hotspots," or where the speaker spends most of their time in the formant space. While traditional VSA exclusively provides information about the perimeter of the speaker's articulatory space, VSD allows for a more nuanced examination of the entire speech signal. Whitfield and Mehta's (2019) examination of VSD in speakers with PD (n = 15) and healthy speakers (n = 15) exemplifies this measure's value. They found that VSD_{10}, or the area within the innermost 90% of the formant density distribution, was comparable between speakers with and without PD (p = .15). In contrast, VSD_{90}, or the area within the innermost 10% of the formant density distribution, was significantly smaller for PD speakers (p = .012). Together, these findings suggest that speakers with PD *can* reach vowel targets in peripheral formant regions like neurologically healthy speakers. However, most of their articulatory–acoustic behavior is confined to a smaller formant space.

Since their proposal, these alternative VSA measures have primarily been investigated for their utility in distinguishing between speakers with and without dysarthria (Sapir et al., 2010; Story & Bunton, 2017; Whitfield et al., 2018; Whitfield & Mehta, 2019). In contrast, little is known about how these alternative measures relate to speech intelligibility, which is one of the clinical strengths of traditional VSA. Before adopting these measures as a replacement for the traditional VSA measure, we must first determine if and to what degree these measures relate to speech intelligibility. Understanding these relationships would provide objective and potentially automatic indices of speech intelligibility.

This Study

This study examined the relationship between speech intelligibility and several measures of vowel space, including

traditional token-based measures, such as VSA and corner dispersion, as well as novel trajectory-based measures, such as vowel space hull area (VSA_{Hull}) and VSD. The study aimed to determine if alternative VSA measures (i.e., the novel trajectory-based measures, VSA_{Hull} and VSD) predicted speech intelligibility to the same extent as the traditional vowel measures (i.e., the token-based measures, VSA and corner dispersion). These vowel measures were obtained from a relatively large sample size consisting of speakers with dysarthria of various etiologies and severities. Based on the previous literature, the strength of the relationship between acoustic measures and intelligibility is expected to vary depending on how intelligibility is measured. Therefore, this study measured intelligibility in two ways, including OTs and visual analog scale (VAS) ratings. The following research questions were posed: (a) How well do trajectory-based and token-based vowel space measures predict intelligibility? And (b) does the relationship between vowel measures and intelligibility differ based on the type of intelligibility measurement (i.e., OTs vs. VAS ratings)?

We hypothesized that traditional VSA and corner dispersion would show a moderate positive relationship to intelligibility based on the previous literature (Y.-J. Kim et al., 2011; Lansford & Liss, 2014a; Weismer et al., 2001). Additionally, because VSD measures obtained from high-density thresholds (e.g., VSD_{90}) are sensitive to the presence of dysarthria (Whitfield & Mehta, 2019), we hypothesized that VSD_{75}, or the area of the innermost 25% of the formant density distribution, would demonstrate a significant positive relationship to intelligibility. To date, VSD measures have been explored in two studies using a number of density thresholds (Story & Bunton, 2017; Whitfield & Mehta, 2019). Story and Bunton (2017) recommended the use of a .25 density threshold, while Whitfield and Mehta (2019) examined VSD at numerous thresholds, ranging from .10 to .90, including threshold levels at every .10 increments (i.e., .10, .20, ..., .90). To limit the number of VSD measures included in the study, we examined VSD_{25} and VSD_{75} to represent the lower and upper threshold levels of the VSD area.

Furthermore, we hypothesized that the type of measure for intelligibility (OT vs. VAS) would impact the relationship between the acoustic measures and intelligibility. Notably, in contrast to several previous studies, this study did not investigate these acoustic-to-perceptual relationships using a DME measure of intelligibility. The decision to investigate VAS ratings instead was motivated by the potential for these findings to be generalized to clinical settings. While DME methods are commonly used within research (Y.-J. Kim et al., 2011; Tjaden & Wilding, 2004; Turner et al., 1995), this method has no documented use within clinical settings. Most speech-language pathologists report measuring intelligibility by estimating the percentage of understood speech following a conversation with their patient (King et al., 2012). We argue that VAS is a commonly used measure in research settings (Chiu & Neel, 2020; Tjaden et al., 2014) that could be generalized to clinical settings as an evidence-based replacement for the currently used percent estimation method (i.e., similar to existing tools using VAS scales, such as the Consensus Auditory-Perceptual Evaluation of Voice [Zraick et al., 2011]).

Method

This study was approved by the Florida State University Institutional Review Board (FSU-IRB STUDY00002322).

Participants

One hundred and forty naïve listeners, ages 18–74 years (M_{age} = 33.23, SD = 13.43), participated in this study. Participants were recruited via the online crowdsourcing platform, Prolific (Palan & Schitter, 2018). To participate in the study, the participants met the following inclusion criteria: (a) 18 years old or older, (b) fluent speakers of English, and (c) located in the United States of America. The sample included 83 women, 52 men, two nonbinary individuals, two agender individuals, and one genderqueer individual. The listeners were primarily White/Caucasian (n = 110), whereas 11 participants were Asian American, eight were Black/African American, eight were biracial or multiracial, one was Native American/African American, and two individuals preferred not to disclose this information. Of the 140 listeners, two listeners spoke English fluently as a second language. Finally, 16 of the participants were Hispanic/Latino, 122 of the participants were not, and two participants chose not to disclose this information. The listeners were compensated $3 for their participation and, on average, took 15.05 min to complete the experiment.

Additionally, data from 40 speakers with dysarthria, ranging in age between 31 and 87 years (M_{age} = 63.03, SD = 13.63), were used in the study. The dysarthria recordings were previously collected as part of a larger study conducted by the Motor Speech Disorders Lab at Arizona State University (Liss et al., 2013). This sample was balanced between speaker sex (i.e., 20 males and 20 females). All speakers were diagnosed with dysarthria secondary to various etiologies, including PD (n = 10), ALS (n = 10), HD (n = 10), and ataxia (n = 10). The race and ethnicity of these participants are unknown.

Procedure

Intelligibility data were collected via Qualtrics (2022). Participants were encouraged to use headphones

and complete the study in a quiet environment. Following a brief demographic questionnaire, the listeners completed two intelligibility measurements, including OTs and VAS ratings of intelligibility. Audio recordings of The Grandfather Passage were used to obtain measures of intelligibility. Each listener was randomly assigned two speakers. For the first speaker, the listener was presented with half of The Grandfather Passage phrase by phrase. Each phrase could only be played once, and listeners were instructed to type what they heard. Listeners were encouraged to guess what they heard when they could not understand the speaker. Following the transcriptions, the listeners rated the intelligibility of that speaker using a horizontally oriented VAS, with anchors labeled *cannot understand anything* and *understand everything* on the left and right sides, respectively. This process was repeated for the second speaker on the other half of The Grandfather Passage to prevent passage familiarization effects. At the completion of the perceptual experiment, each speaker was rated by four to 12 listeners ($M = 7$, $SD = 2.22$).

Measures

Perceptual Measures

Two intelligibility measures were collected from the naïve listeners, including OTs and VAS ratings of intelligibility. For the OT measure, intelligibility was calculated as the percentage of words correctly transcribed. Transcription accuracy was computed using the *AutoScore* package in R (Borrie et al., 2019). The transcription accuracy scores across phrases and listeners were averaged to obtain a mean transcription accuracy score per speaker. For the VAS ratings of intelligibility, the same process was used to obtain a mean VAS intelligibility rating per speaker.

Acoustic Measures

Five acoustic measures were examined in the present study, including two manual token-based measures (i.e., VSA and corner dispersion) and three automatic trajectory-based measures (i.e., VSA$_{Hull}$, VSD$_{25}$, and VSD$_{75}$). Figure 1 contains examples of the selected acoustic measures, which are described in greater detail below. The code used to treat the data and calculate the target token- and trajectory-based measures can be found in the supplemental information (https://osf.io/hr7aj/).

Data cleaning and preparation. Two separate data-cleaning and preparation protocols were employed, one for the manual token-based measures and one for the automatic trajectory-based measures. For the token-based measures (i.e., VSA and corner dispersion), tokens were manually segmented and extracted from the passage using the *To TextGrid* function in Praat (Boersma & Weenink, 2012), and the F1 and F2 formant values were visually inspected and formant tracking errors manually corrected

Figure 1. Token-based (top row) and trajectory-based (bottom row) measures used in the current study. Although the token-based measures (top row) did not rely on the continuously sampled formant trajectories, they are plotted behind the token measures to demonstrate that all the target measures were measured from the same passage. VSA = vowel space area; VSA$_{Hull}$ = vowel space hull area; VSD = vowel space density; F1 = first formant fequency; F2 = second formant frequency.
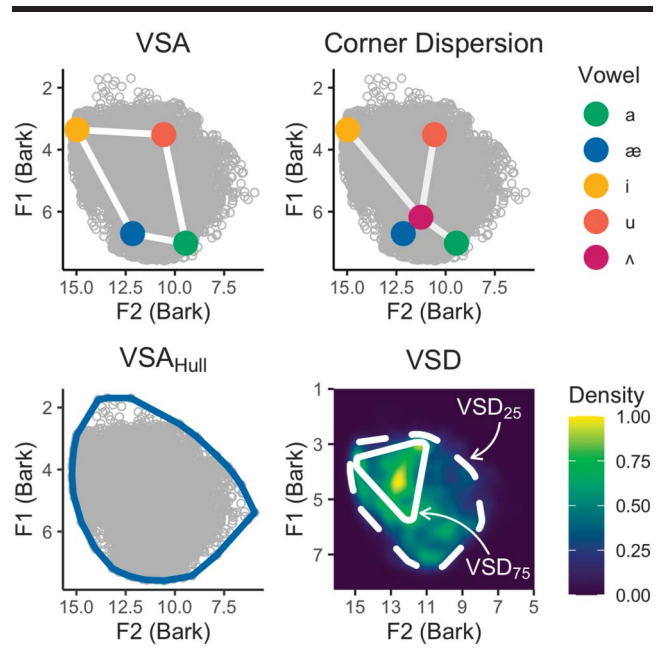
using TF32 (Milenkovic, 2004). Because the speaker sample contained both males and females, the formant data were Bark-transformed to account for sex differences. We acknowledge that this is an imperfect method for normalizing the speakers' sex differences (Clopper, 2009). However, we believe that sex effects will not greatly impact the findings with a balanced sample between males and females. Within the supplemental information, we have included a table containing the summary statistics for the target measures in hertz to allow comparisons to previous studies using hertz-based measures (located at https://osf.io/hr7aj/).

For the trajectory-based measures (i.e., VSA$_{Hull}$, VSD$_{25}$, and VSD$_{75}$), formant frequency data were extracted from the recordings using Praat's *To Formant (burg)* function. Formants were obtained at a time step of .01 s, with a max number of five formants, a window length of .05 s, a pre-emphasis from 50 Hz, and a formant ceiling of 5500 Hz for female speakers and 5000 Hz for male speakers. This process yielded a set of raw formant values for the entire The Grandfather Passage (see Figure 2, Pane A). However, in this study, we were only interested in the formant values for vowel data. Therefore, a three-step data-cleaning process was employed to remove the noisy aperiodic acoustic data created by consonant data and general formant tracking errors (Whitfield & Mehta, 2019). First, formant data during

Figure 2. Formant data treatment. F1 = first formant frequency; F2 = second formant frequency.
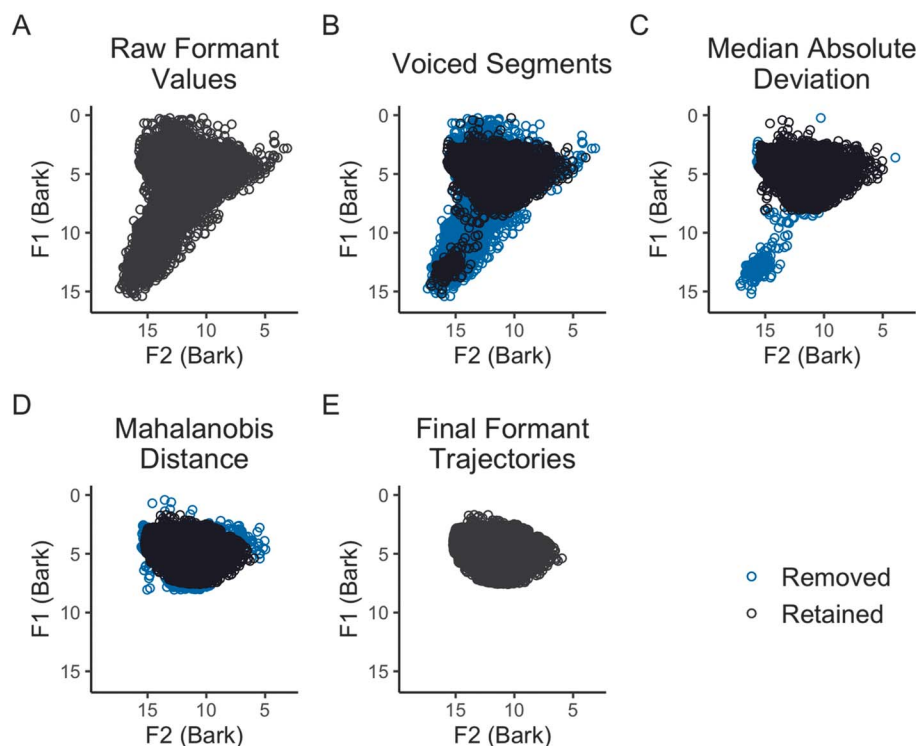
non-voiced segments were removed from the data by omitting data with missing pitch data (see Figure 2, Pane B). Second, the median absolute deviation was calculated using the *mad* function within the *stats* R package (R Core Team, 2022). Coordinate pairs with a median absolute deviation value of 2.5 *SD*s above or below the median were removed from the data (see Figure 2, Pane C). Finally, the multivariate outliers for F1 and F2 were identified by calculating the Mahalanobis distance using the *mahalanobis* function within the *stats* R package (R Core Team, 2022). Coordinate pairs with a Mahalanobis distance 2 *SD*s above or below the mean were removed from the data (see Figure 2, Pane D). Finally, like the token-based measure formant data, the formant frequency data for the automatic trajectory-based measures were Bark-transformed to account for sex differences.

VSA (Bark²). The VSA measure was calculated as the planar area among the Bark-transformed F1–F2 coordinate pairs for the corner vowels /i/, /æ/, /u/, and /a/ (Kent & Y-J. Kim, 2003). The F1–F2 data were obtained from the manually segmented vowel formant data. Before calculating VSA, several tokens of each vowel were identified from the passage, and the F1–F2 values obtained from the temporal midpoint of each token were averaged together to obtain a single set of F1–F2 values per vowel. Finally, the *polyarea* function within the *geometry* R package was used to calculate VSA among the four corner vowels (Roussel et al., 2022).

Corner dispersion (Bark). Corner dispersion was calculated as the Euclidean distance between the Bark-transformed F1–F2 coordinate pairs of the center vowel /ʌ/ and each corner vowel (Lansford & Liss, 2014a; Tjaden et al., 2013; Turner et al., 1995). Again, the F1–F2 data were obtained from the temporal midpoint of the manually segmented vowels. This process yielded four distance measures per speaker (i.e., the distance between /ʌ/–/i/, /ʌ/–/æ/, /ʌ/–/u/, and /ʌ/–/a/). These four distance measures were averaged to obtain a single corner dispersion measure for each speaker.

VSA$_{Hull}$ (Bark²). VSA$_{Hull}$ is calculated as the convex hull area of the Bark-transformed passage-level formant data (Whitfield et al., 2018; Whitfield & Mehta, 2019). This measure was calculated using the automatically extracted and treated formant data. The convex hull area was computed using the *chull* function within the *grDevices* R package (R Core Team, 2022).

VSD (Bark²). VSD was calculated as the convex hull area of the Bark-transformed passage-level formant data at the normalized thresholds of .25 and .75 (i.e., the vowel space's 75% and 25% innermost regions, respectively). As previously mentioned, a few studies have investigated the VSD measure using varying methodologies for its calculation. This study used the methods outlined in Story and Bunton (2017). First, the Bark-transformed data were divided into a grid containing 250 bins using the *kde*

function in the *ks* R package (Duong, 2007). This process yielded a density estimate for each of the 250 bins, which were then re-expressed to a range of 0–1 using the *rescale* function in the *scales* R package (Wickham et al., 2016). Finally, subsets of the density data (i.e., data with density values greater than .25 and .75) were used to calculate the convex hull area using the method described for the VSA$_{Hull}$ measure. This process yielded two measures of convex hull area at the .25 and .75 density thresholds, hereafter referred to as VSD$_{25}$ and VSD$_{75}$, respectively.

Reliability

Acoustic segmentation for the token measures was completed by the first and second authors. To ensure inter-measurer reliability, 20% of the data (eight of the 40 speakers) were segmented by both examiners. The reliability between the two sets of measurements was examined using single and average intraclass correlation coefficients (ICCs). For the F1 midpoint values, the two sets of measurements yielded a single ICC score of .87 with a 95% confidence interval (CI) from .84 to .89, $F(359, 348) = 14$, $p < .001$. The average ICC score was .93 with a 95% CI from .91 to .94, $F(359, 347) = 14$, $p < .001$. For the F2 midpoint values, the two sets of measurements yielded a single ICC score of .93 with a 95% CI from .92 to .94, $F(358, 359) = 28$, $p < .001$. The average ICC score was .96 with a 95% confidence interval from .96 to .97, $F(358, 359) = 28$, $p < .001$. Thus, the interrater reliability for the acoustic segmentation was considered good to excellent.

Statistical Analysis

For our first research question, a hierarchical regression approach was used to examine the relationship between the vowel acoustic measures and intelligibility. Hierarchical regression is similar to stepwise regression in that it examines the effects of predictors as they are sequentially added to the model. However, unlike stepwise regression, the order in which predictors are entered into the analysis is informed by theory for the hierarchical regression method (Berger, 2004). This approach was selected for the current analysis due to expected issues with multicollinearity among the vowel measures. The model creation process was completed in the following steps: (a) create an initial model with one predictor, (b) add an additional predictor for a second model, (c) compare the fit of the initial model with the second model, (d) remove variables with variance inflation factor (VIF) values greater than 5, (e) continue this process until all five acoustic measures have been tested, (f) select the best fitting model, and (g) remove nonsignificant predictors to obtain the most parsimonious model. This process was completed twice, modeling OTs and VAS ratings of intelligibility as the outcomes.

The models were built in R using the *lm* function from the *stats* package (R Core Team, 2022), and the model assumptions were checked using the *check_model* function from the *performance* package (Lüdecke et al., 2021). Multicollinearity was examined using VIF values, with VIF values greater than 5 indicating that multicollinearity issues were present. Finally, to compare the fit of the nested models, the *anova* function of the *stats* package was used. Prior to model creation, the relationships among the measures of interest were examined with a correlation matrix using the *rcorr* function from the *Hmisc* package (Harrell, 2022).

The effect of the intelligibility method (i.e., OT and VAS) on the relationship between the acoustic measures and intelligibility was investigated by visually inspecting the hierarchical regression model scatter plots created for our first research question. Also, the effect of the intelligibility methods was examined by comparing the R^2 values of the OT and VAS models. Finally, for our second research question, which directly examined the relationship between OT and VAS methods, a linear regression model was created using the *lm* function in the *stats* package (R Core Team, 2022). The OT scores served as the outcome measure, with VSA, speaker sex, and speaker etiology as the predictors. An adjusted alpha level of $\alpha_{Bonferroni} = .016$ was used to evaluate the significance of the models created for our two research questions.

Results

The cleaned data, analysis code, and output are provided publicly at https://osf.io/hr7aj/. Table 2 contains the summary statistics for the five vowel acoustic measures and the two intelligibility measures across etiology and sex.

The correlations matrix for the acoustic and perceptual measures is presented in Table 3. As expected, there were some strong correlations among the five vowel measures. However, the strength of these relationships varied by the type of vowel measure, such that measures were correlated more strongly to other measures of the same type. In other words, manual token-based measures showed a strong correlation to each other (i.e., VSA and corner dispersion were strongly correlated, $r = .727$), whereas the trajectory-based measures showed strong relationships with each other (e.g., VSA$_{Hull}$ and VSD$_{25}$ were strongly correlated, $r = .838$). The correlation across token- and trajectory-based measures varied between .13 and .55. Notably, all of the trajectory-based measures were weakly correlated to either intelligibility measure. In contrast, both token-based measures were moderately correlated to both intelligibility measures.

Table 2. Summary statistics for the target acoustic and intelligibility measures.

Variable	All etiologies		ALS		Ataxia		HD		PD	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
All speakers										
VSA (Bark^2)	4.42	2.07	3.32	1.15	5.42	2.45	5.16	1.76	3.79	2.16
Corner disp (Bark)	2.04	0.39	1.87	0.48	2.24	0.38	2.09	0.34	1.98	0.31
VSA$_{Hull}$ (Bark^2)	30.95	8.78	26.37	7.95	33.99	5.66	34.29	8.45	29.16	10.84
VSD$_{25}$ (Bark^2)	15.85	6.74	15.03	7.44	18.64	5.50	16.08	7.03	13.66	6.85
VSD$_{75}$ (Bark^2)	2.26	2.26	2.71	3.07	2.92	2.33	2.18	2.08	1.22	0.98
Intelligibility (VAS)	51.95	26.27	40.76	26.16	54.08	20.84	48.53	31.45	64.44	23.49
Intelligibility (OT)	58.47	23.91	47.45	23.70	62.02	20.32	56.95	27.58	67.46	22.39
Female speakers										
VSA (Bark^2)	5.11	2.52	3.46	0.70	6.46	3.11	6.08	2.07	4.44	2.89
Corner disp (Bark)	2.16	0.41	1.89	0.46	2.48	0.22	2.18	0.39	2.10	0.38
VSA$_{Hull}$ (Bark^2)	35.34	8.58	29.30	9.64	36.00	6.62	39.60	7.30	36.46	9.47
VSD$_{25}$ (Bark^2)	19.45	6.50	19.10	8.55	20.85	5.91	20.15	6.04	17.68	6.99
VSD$_{75}$ (Bark^2)	2.80	2.49	3.46	3.97	3.20	2.27	2.75	2.34	1.80	1.07
Intelligibility (VAS)	48.41	31.31	27.98	19.70	51.33	28.35	50.54	37.95	63.81	34.49
Intelligibility (OT)	55.25	28.98	39.42	21.25	59.19	27.98	56.49	34.75	65.90	32.72
Male speakers										
VSA (Bark^2)	3.73	1.20	3.17	1.55	4.38	1.04	4.23	0.77	3.14	1.05
Corner disp (Bark)	1.92	0.35	1.84	0.56	2.00	0.36	2.00	0.29	1.85	0.15
VSA$_{Hull}$ (Bark^2)	26.57	6.63	23.43	5.28	31.99	4.27	28.99	6.08	21.86	6.44
VSD$_{25}$ (Bark^2)	12.26	4.88	10.95	3.12	16.44	4.59	12.02	5.77	9.65	4.05
VSD$_{75}$ (Bark^2)	1.71	1.90	1.97	2.02	2.64	2.62	1.61	1.85	0.64	0.41
Intelligibility (VAS)	55.49	20.24	53.54	27.26	56.83	12.44	46.52	27.83	65.06	7.13
Intelligibility (OT)	61.69	17.67	55.48	25.52	64.85	11.23	57.42	22.43	69.01	7.17

Note. ALS = amyotrophic lateral sclerosis; Ataxia = cerebellar ataxia; HD = Huntington's disease; PD = Parkinson's disease; VSA = vowel space area; Corner disp = corner dispersion; VSAHull = vowel space hull area; VSD = vowel space density; VAS = visual analog scale; OT = orthographic transcription.

Research Question 1

The hierarchical regression model results for predicting OT accuracy are reported in Table 4. This table shows the sequential order in which the predictors were entered into the model. The novel trajectory-based measures were entered first (Models 1–3), followed by the traditional token-based measures (Models 4 and 5). In Model 3, the VSD$_{25}$ measure was removed from the model due to a high VIF value (i.e., VIF > 5). The results revealed that traditional VSA was the only significant predictor of OT accuracy and was, therefore, the sole predictor retained in the final, most parsimonious model, $t(38) = 3.64$, $p <$ .001, 95% CI [2.60, 9.14]. The R^2 was .259, suggesting

that the variance of VSA could explain 25.9% of the variance in the OT accuracy.

The hierarchical regression model results for predicting VAS ratings of speech intelligibility are reported in Table 5. Like the OT model, the novel trajectory-based measures were entered into the model first, followed by the token-based measures. Again, in Model 3, the VSD$_{25}$ measure was removed from the model due to a high VIF value (i.e., VIF > 5). Like the OT model results, traditional VSA was the best predictor of VAS ratings of intelligibility and was the sole predictor retained in the final model, $t(38) = 3.43$, $p = .001$, 95% CI [2.52, 9.80]. The R^2 was .236, suggesting that the variance of VSA could explain 23.6% of the variance in the VAS ratings of intelligibility.

Table 3. Correlation matrix for the target acoustic measures and the two intelligibility measures.

Measure	VSA	Corner disp	VSA$_{Hull}$	VSD$_{25}$	VSD$_{75}$	Intelligibility (VAS)
VSA						
Corner disp	.73					
VSA$_{Hull}$	.55	.49				
VSD$_{25}$	.52	.40	.84			
VSD$_{75}$	.28	.13	.46	.68		
Intelligibility (VAS)	.49	.40	.25	.15	.09	
Intelligibility (OT)	.51	.42	.29	.18	.09	.95

Note. VSA = vowel space area; Corner disp = corner dispersion; VSA$_{Hull}$ = vowel space hull area; VSD = vowel space density; VAS = visual analog scale ratings of intelligibility; OT = orthographic transcription accuracy.

Table 4. Hierarchical regression results for the models predicting orthographic transcription accuracy.

Predictors	Model 1 Estimates	Model 2 Estimates	Model 3 Estimates	Model 4 Estimates	Model 5 Estimates	Final model Estimates
(Intercept)	48.30**	47.34***	33.23*	30.77*	24.43	32.51***
VSD$_{25}$ (Bark^2)	0.64	0.80	—	—	—	—
VSD$_{75}$ (Bark^2)		−0.72	−0.62	−0.82	−0.58	—
VSA$_{Hull}$ (Bark^2)			.86	0.12	—	—
VSA (Bark^2)				5.84**	5.22*	5.87***
Corner disp (Bark)					6.00	—
R^2/R^2 adjusted	.033/.007	.035/−.017	.086/.037	.263/.202	.267/.206	.259/.239

Note. Em dashes indicate variables that were removed from the model. VSD = vowel space density; VSA$_{Hull}$ = vowel space hull area; VSA = vowel space area; Corner disp = corner dispersion.

*$p < .05$. **$p < .01$. ***$p < .001$.

Corner dispersion is likely not a significant predictor in the hierarchical regression models because of its strong relationship to VSA ($r = .727$). However, despite this strong relationship, corner dispersion was included in models because it did not violate our predetermined criteria for identifying multicollinearity (i.e., VIF > 5). However, to explore the relationship between corner dispersion and intelligibility, two simple linear regression models were created with corner dispersion as the sole predictor. The results indicated that corner dispersion was a significant predictor of intelligibility when examined in isolation, $t(38) = 2.86$, $p = .007$, for OT and $t(38) = 2.70$, $p = .01$, for VAS. However, with R^2 values of .18 and .16 for the OT and VAS models, respectively, the variance of corner dispersion accounted for less of the variance in intelligibility compared with the VSA models (i.e., .26 and .24 for OT and VAS, respectively).

Research Question 2

The effect of the intelligibility method (i.e., OT and VAS) on the relationship between the acoustic measures and intelligibility can be seen in the results for the first research question. Specifically, the model results were the same, regardless of how intelligibility was measured.

Furthermore, Figure 3 shows the effect of the intelligibility method on the relationship between the acoustic measures and intelligibility. Despite differences in individual responses, the relationship or slope between intelligibility and the various acoustic measures was very similar between the OT and VAS methods. However, VAS appears to consistently underestimate the OT scores, as indicated by lower intercepts observed for all measures.

The linear regression model predicting OTs using VAS ratings of intelligibility is reported in Table 6. In the first model, VAS, speaker etiology, speaker sex, and the interactions between VAS ratings and etiology and between VAS ratings and sex were entered into the model. However, VAS was the only significant predictor and, therefore, was the only predictor retained in the final model. Theoretically, a perfect relationship between OTs and VAS ratings of intelligibility would yield an intercept of 0 and a VAS slope of 1. In the final model, the intercept estimate was significantly different from 0, $t(38) = 4.93$, $p < .001$, 95% CI [8.10, 19.39]. Additionally, the slope for VAS was significant, $t(38) = 17.94$, $p < .001$, 95% CI [0.76, 0.96]. Together, these results indicate a strong positive relationship between OTs and VAS ratings of intelligibility. However, the significant intercept indicates that VAS ratings underestimated

Table 5. Hierarchical regression results for the models predicting visual analog scale ratings of intelligibility.

Predictors	Model 1 Estimates	Model 2 Estimates	Model 3 Estimates	Model 4 Estimates	Model 5 Estimates	Final model Estimates
(Intercept)	42.63***	42.34***	25.04	25.22	16.61	24.70**
VSD$_{25}$ (Bark^2)	0.59	0.64	−1.12	—	—	—
VSD$_{75}$ (Bark^2)		−0.22	0.88	−0.58	−0.41	—
VSA$_{Hull}$ (Bark^2)			1.38	0.00	−0.05	—
VSA (Bark^2)				6.34**	5.49	6.16**
Corner disp (Bark)					6.69	
R^2/R^2 adjusted	.023/−.003	.023/−.030	.081/.004	.238/.175	.243/.156	.236/.216

Note. Em dashes indicate variables that were removed from the model. VSD = vowel space density; VSA$_{Hull}$ = vowel space hull area; VSA = vowel space area; Corner disp = corner dispersion.

**$p < .01$. ***$p < .001$.

Figure 3. Linear relationships between the selected acoustic measures and intelligibility. Intelligibility was measured using orthographic transcription (OT) accuracy (the solid black line) and visual analog scale (VAS) ratings of intelligibility (the dashed blue line). VSA = vowel space area; VSA_Hull = vowel space hull area; VSD = vowel space density.
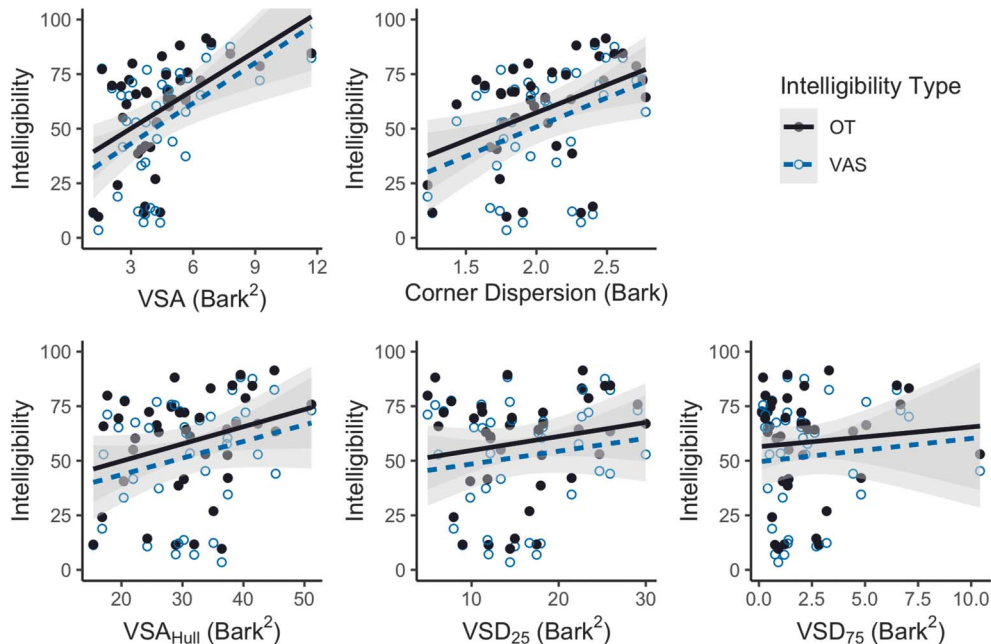
intelligibility compared to OTs. These results can be observed in Figure 4.

Exploratory Analysis

In addition to the analyses described in the methods, additional post hoc exploratory analyses were completed to examine how speaker etiology (i.e., ataxic, ALS, PD,

Table 6. Linear regression models predicting orthographic transcriptions of intelligibility.

Predictors	Model 1 Estimates	Final model Estimates
(Intercept)	10.90	13.75***
Etiology[ALS]	Reference	
Etiology[Ataxia]	2.99	
Etiology[HD]	2.10	
Etiology[PD]	−2.49	
Sex[F]	Reference	
Sex[M]	6.89	
VAS	0.89***	0.86***
VAS:Etiology[Ataxia]	−0.00	
VAS:Etiology[HD]	0.00	
VAS:Etiology[PD]	0.03	
VAS:Sex[M]	−0.12	
R^2/R^2 adjusted	.903/.874	.894/.892

Note. ALS = amyotrophic lateral sclerosis; Ataxia = cerebellar ataxia; HD = Huntington's disease; PD = Parkinson's disease; F = female; M = male; VAS = visual analog scale.

***$p < .001$.

and HD) may affect the predictive relationship between the selected acoustic measures and intelligibility. The speaker's etiology was not included in our original analysis for two reasons. First, this study aimed to replicate the findings of previous studies that have reported mild-to-moderate relationships between VSA and intelligibility ($R^2 = .16–.46$), all of which examined this relationship across many dysarthria etiologies, such as speakers with PD, stroke, traumatic brain injury, and multiple systems atrophy (Y-J. Kim et al., 2011); speakers with PD, ALS,

Figure 4. Relationship between orthographic transcriptions (OTs) and visual analog scale (VAS) ratings of intelligibility. This figure displays individual speaker data and the linear trend lines for the four etiology groups. The solid black line signifies the perfect theoretical slope. Data to the left of the reference line indicate that VAS ratings underestimated OT scores, whereas data to the right of the line indicate that VAS ratings overestimated OT scores. ALS = amyotrophic lateral sclerosis; HD = Huntington's disease; PD = Parkinson's disease.
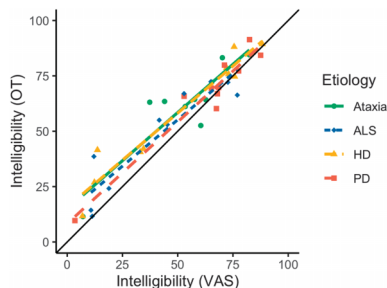
ataxia, and HD (Lansford & Liss, 2014a); and speakers with PD and ALS (Weismer et al., 2001). Second, with only 10 speakers within each etiology group, this study was not sufficiently powered to statistically examine and make conclusions about the effects of etiology on the relationship between intelligibility and the selected acoustic measures. Therefore, we opted to explore these relationships descriptively using data visualization.

Figure 5 displays the effect of etiology on the relationship between intelligibility and the five acoustic measures of interest. With such few data points for each etiology group, the etiology-specific relationship between the acoustic measures and intelligibility is highly sensitive to outliers within the data. Take, for example, the ALS group within the VSD_{75} facet. The relationship between intelligibility and VSD_{75} is heavily influenced by a single speaker with a large VSD_{75}. Therefore, it is challenging to make conclusions about the nature of these relationships with the current data alone. Notably, among these five acoustic measures, the trend lines for the VSA measure are generally similar across the four etiology groups. That is, as VSA increased, speakers were perceived to be more intelligible. With a larger sample size, future work should focus on examining the etiology-specific relationship between intelligibility and acoustic measures of vowel space.

Discussion

This study examined the relationship between speech intelligibility and various acoustic measures of vowel space, including traditional token-based measures and more novel trajectory-based measures of vowel space. This study also examined how these relationships are impacted by the method used to measure intelligibility, including OTs and VAS ratings of speech intelligibility. In summary, only the token-based measures were significant predictors of intelligibility (i.e., traditional VSA and corner dispersion), and traditional VSA was the best overall predictor of intelligibility. Additionally, VAS ratings of intelligibility were found to underestimate the OT scores. However, the intelligibility method did not significantly impact the relationship between intelligibility and the various acoustic measures.

Acoustic-to-Perceptual Relations

Of all the measures, traditional VSA demonstrated the strongest correlation to intelligibility (OT: $r = .51$; VAS: $r = .49$). Furthermore, the final models for predicting OTs and VAS ratings of intelligibility included traditional VSA as the sole predictor. The R^2 values for these models were .26 and .24, suggesting a moderate and positive relationship between VSA and intelligibility. This finding is consistent with previous work that has examined the relationship between VSA and intelligibility with relatively large samples and across a wide range of dysarthria severities and etiologies (Y-J. Kim et al., 2011; Lansford & Liss, 2014a; Weismer et al., 2001). This study adds to the growing evidence of a moderate and positive relationship between VSA and intelligibility, especially when speakers of various etiologies and dysarthria severities are considered.

Like traditional VSA, corner dispersion was significantly related to intelligibility and demonstrated a moderate correlation to both OT and VAS (OT: $r = .42$; VAS: $r = .40$). Although corner dispersion was not retained in the final models created for our first research question, corner dispersion was found to be a significant predictor of intelligibility when examined as the sole predictor of

Figure 5. The effects of etiology on the relationship between intelligibility and the acoustic measures. VSA = vowel space area; VSA_{Hull} = convex hull area; $VSD_{25, 75}$ = vowel space density at the .25 and .75 density thresholds; OT = orthographic transcription; VAS = visual analog scale ratings of intelligibility; ALS = amyotrophic lateral sclerosis; HD = Huntington's disease; PD = Parkinson's disease.
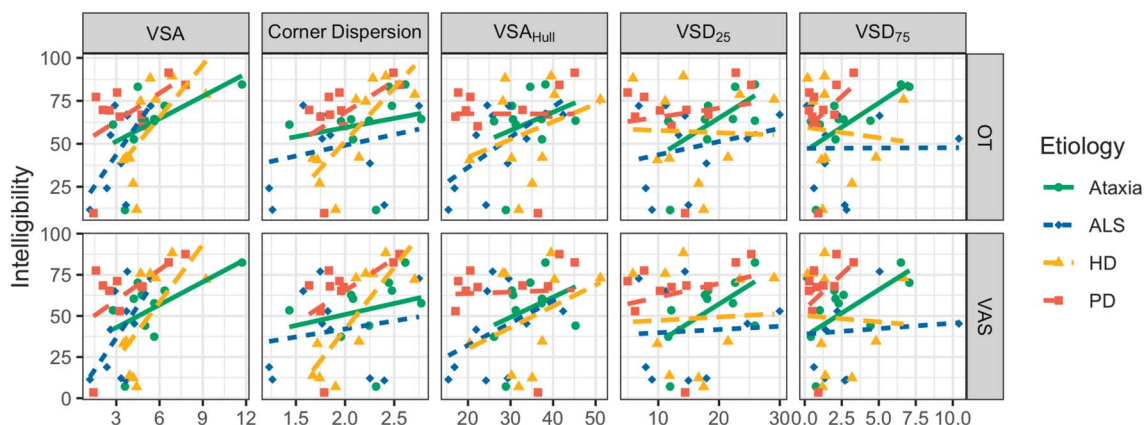
intelligibility. However, by comparing the R^2 values for the sets of models, it was determined that VSA was a better predictor of intelligibility. These findings are not surprising, given that corner dispersion and traditional VSA were highly correlated ($r = .73$). These findings indicate that, despite being labor intensive measures derived from the manually segmented corner vowels are sensitive to intelligibility changes and may serve as an objective measure of dysarthria severity.

It is essential to note that speech intelligibility is an intricate combination of several speech domains, including the articulatory, phonatory, resonatory, and respiratory subsystems. Therefore, we cannot expect one measure to serve as a proxy for speech intelligibility. Instead, a successful model to explain speech intelligibility would likely contain various measures representing a span of speech subsystems. However, before developing an explanatory model of intelligibility, we need to understand which measures show promise for predicting speech intelligibility. The current findings suggest that both traditional VSA and corner dispersion may show promise for explaining at least part of the variance of speech intelligibility.

Surprisingly, none of the trajectory-based measures were significant predictors of intelligibility. One possible explanation for this finding is that global, trajectory-level measures are indeed not sensitive to intelligibility. Instead, it may be the case that only specific segments of speech (i.e., the corner vowels) are sensitive to intelligibility. This explanation is supported by previous literature documenting specific phonetic contexts to be more sensitive to the presence or severity of dysarthria than others (Y-J. Kim, 2017; Y-J. Kim et al., 2009; Rosen et al., 2008). For example, Y-J. Kim et al. (2009) investigated the relationship between F2 slope and speech intelligibility in speakers with dysarthria due to PD and stroke and neurologically healthy control speakers. This relationship was examined separately for six different words. The results indicated that F2 slope was significantly related to speech intelligibility only for the words "shoot" and "wax" ($R^2 = .143$ for "shoot" and $R^2 = .139$ for "wax"). This study and others

highlight the value of a segment-specific approach to investigating intelligibility deficits in dysarthria. Therefore, within this study, it may be the case that measures derived from the corner vowels may be sensitive to intelligibility, whereas global measures derived from the entire formant space are not.

Another potential explanation for the nonsignificant relationship between intelligibility and trajectory-based measures is the methods used to clean and treat the data. The trajectory-based measures examined within this study are derived from the formant trajectory space obtained from passage reading. Before these measures can be calculated, the formant space must be treated to obtain data from just the voiced vowel formants that do not contain any formant tracking errors (as seen in Figure 2). However, there is no standard method for filtering and cleaning formant frequency data, resulting in varying methods from study to study (Sandoval et al., 2013; Whitfield & Mehta, 2019).

To our knowledge, there has been no attempt to empirically investigate the best practices for signal processing of the speech signal. Therefore, it will be challenging to compare the same measures across different studies until this effort is made. In this study, we applied a three-step data-cleaning process that has been employed in previous studies (Whitfield & Mehta, 2019). This process includes (a) removing non-voiced formant data, (b) removing univariate outliers using a median absolute deviation function and a cutoff of 2.5 SDs from the mean, and (c) removing multivariate outliers using Mahalanobis distance and a cutoff of 2 SDs from the mean. However, there is no evidence to support the use of these specific techniques or cutoff values. Figure 6 demonstrates VSA$_{Hull}$'s sensitivity to varying cutoff values for the Mahalanobis distance outlier detection. Based on the cutoff values for Mahalanobis distance alone, the VSA$_{Hull}$ area can vary approximately 14 Bark^2, which is not a negligible difference.

While the trajectory-based measures in this study did not demonstrate a meaningful relationship to speech

Figure 6. Vowel space area hull at various Mahalanobis distance cutoff points. F1 = first formant frequency; F2 = second formant frequency.
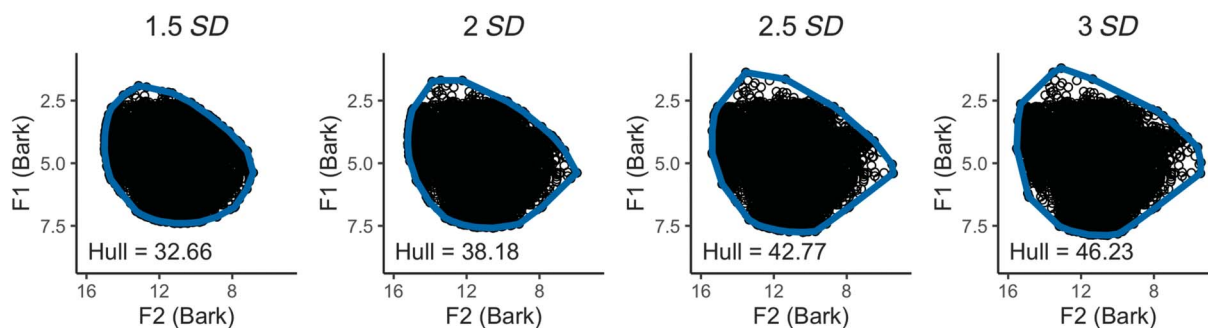
intelligibility, these findings do not undermine the value of these measures. As previously mentioned, these measures are evidenced to detect (a) group differences between speakers with and without dysarthria (Whitfield & Mehta, 2019) and (b) changes during various speaking conditions, such as loud speech (Whitfield et al., 2018). Additionally, it is possible that these trajectory-based measures are related to perceptual outcomes other than intelligibility, such as judgments of articulatory precision, naturalness, or dysarthria severity. More research examining the acoustic-to-perceptual relations for these measures is needed.

Methodological Considerations for Measuring Intelligibility

The method used for measuring intelligibility (i.e., either VAS or OT) does not significantly affect the relationship between the vowel space measures and intelligibility. This finding suggests that, for the purpose of examining the relationship between objective measures and intelligibility, VAS may be used to measure intelligibility instead of the labor-intensive OT method. Notably, VAS ratings of intelligibility tended to underestimate the OT scores across the spectrum of intelligibility by approximately 13.75 points (i.e., we are 95% confident that the mean value that VAS underestimated OT scores is between 8.10 and 19.39). This finding supports previous research that has observed VAS ratings to underestimate OT scores, despite a moderate-to-strong correlation between the two measures (Stipancic et al., 2016).

Limitations and Future Directions

This study is not without limitations. First, as previously mentioned, the formant data were treated using one of many existing protocols for signal processing of the formant data. Future work should strive to establish an evidence-based approach for data cleaning and signal processing. Additionally, this study only utilizes VSD calculated at only the .25 and .75 density thresholds. However, previous research has included various VSD threshold levels (i.e., VSD_{10}–VSD_{90}; Whitfield & Mehta, 2019). Future work should examine how VSD relates to intelligibility at different threshold levels.

Furthermore, this study only examined dysarthria from four etiology groups and did not include a neurologically healthy control speaker group. While the objective of this study was not to examine group differences, future work should explore these relationships in other clinical and nonclinical populations to determine if the current findings generalize to other populations. Additionally, due to the small sample size within each etiology group (n =

10), we could not statistically examine the effect of the etiology on the relationship between intelligibility and the various acoustic vowel space measures. Future work in this research line should examine the effect of etiology with a larger subgroup sample size.

Another limitation of this study is that we utilized prerecorded samples of The Grandfather Passage to obtain OTs and VAS ratings of intelligibility from naïve listeners. While we employed some efforts to reduce passage familiarization, the passage context may have facilitated the listeners' transcriptions by providing semantic and syntactic cues. Future work could derive OT measures from semantically anomalous phrases to avoid potential top-down influence. Finally, future research should examine how these acoustic-to-perceptual relations may or may not differ when intelligibility is measured using DME.

Conclusions

In conclusion, this study examined the relationship between speech intelligibility and a handful of acoustic measures of vowel space, both token based and trajectory based. The purpose was to determine if alternative vowel space measures (i.e., the novel trajectory-based measures, VSA_{Hull} and VSD) predicted speech intelligibility to the same extent as the traditional token-based vowel measures (i.e., traditional VSA and corner dispersion). The findings revealed token-based measures as the only significant predictors of intelligibility, with the traditional VSA measure being the strongest predictor of intelligibility. In contrast, the trajectory-based measures (i.e., VSA_{Hull}, VSD_{25}, and VSD_{75}) were not predictive of speech intelligibility. Therefore, we recommend that future work that aims to build explanatory models of speech intelligibility utilize the traditional VSA measure instead of alternative VSA measures (i.e., VSA_{Hull}, VSD_{25}, and VSD_{75}) to characterize the articulatory features of speech. However, this recommendation is based on the current methods used to clean and treat the formant data. It is possible that, with improved methods for the automatic extraction of formant frequencies, these automatic trajectory-based measures may become more effective at predicting speech intelligibility. However, more work is needed to determine how to improve these methods.

Additionally, the method used to measure intelligibility (i.e., OT and VAS) and its impact on the acoustic-to-perceptual relationships were examined. The results revealed that VAS ratings underestimated the OT scores. However, the intelligibility method did not affect the relationship between the vowel space measures and intelligibility. Therefore, these findings suggest that the method used to measure intelligibility (i.e., OT or VAS) does not

meaningfully affect the relationship between acoustic measures of articulation and speech intelligibility.

Data Availability Statement

The cleaned data, analysis code, and output are provided publicly at https://osf.io/hr7aj/. This OSF also contains additional alternative analyses that were completed but not reported in great detail in this article.

References

Berger, D. E. (2004). Using regression analysis. In J. S. Wholey, H. P. Hatry, & K. E. Newcomer (Eds.), *Handbook of practical program evaluation* (2nd ed., pp. 479–505). Jossey-Bass.

Boersma, W., & Weenink, D. (2012). *Praat: Doing phonetics by computer* (Version 6.2.10) [Computer software]. http://www.praat.org/

Borrie, S. A., Barrett, T. S., & Yoho, S. E. (2019). Autoscore: An open-source automated tool for scoring listener perception of speech. *The Journal of the Acoustical Society of America, 145*(1), 392–399. https://doi.org/10.1121/1.5087276

Chiu, Y.-F., & Neel, A. (2020). Predicting intelligibility deficits in Parkinson's disease with perceptual speech ratings. *Journal of Speech, Language, and Hearing Research, 63*(2), 433–443. https://doi.org/10.1044/2019_JSLHR-19-00134

Clopper, C. G. (2009). Computational methods for normalizing acoustic vowel data for talker differences. *Language and Linguistics Compass, 3*(6), 1430–1442. https://doi.org/10.1111/j.1749-818X.2009.00165.x

DuHadway, C. M., & Hustad, K. C. (2012). Contributors to intelligibility in preschool-aged children with cerebral palsy. *Journal of Medical Speech-Language Pathology, 20*(4), 11.

Duong, T. (2007). ks: Kernel density estimation and kernel discriminant analysis for multivariate data in R. *Journal of Statistical Software, 21*(7), 1–16. https://doi.org/10.18637/jss.v021.i07

Fox, R. A., & Jacewicz, E. (2017). Reconceptualizing the vowel space in analyzing regional dialect variation and sound change in American English. *The Journal of the Acoustical Society of America, 142*(1), 444–459. https://doi.org/10.1121/1.4991021

Harrell, F. E., Jr. (2022). *Hmisc: Harrell miscellaneous* (Version 4.7-0) [R package]. https://cran.r-project.org/web/packages/Hmisc/index.html

Higgins, C. M., & Hodge, M. M. (2002). Vowel area and intelligibility in children with and without dysarthria. *Journal of Medical Speech-Language Pathology, 10*(4), 271–277.

Kent, R. D., & Kim, Y-J. (2003). Toward an acoustic typology of motor speech disorders. *Clinical Linguistics & Phonetics, 17*(6), 427–445. https://doi.org/10.1080/0269920031000086248

Kim, H., Hasegawa-Johnson, M., & Perlman, A. (2011). Vowel contrast and speech intelligibility in dysarthria. *Folia Phoniatrica et Logopaedica, 63*(4), 187–194. https://doi.org/10.1159/000318881

Kim, Y-J. (2017). Acoustic characteristics of fricatives /s/ and / ʃ / produced by speakers with Parkinson's disease. *Clinical Archives of Communication Disorders, 2*(1), 7–14. https://doi.org/10.21849/cacd.2016.00080

Kim, Y-J., Kent, R. D., & Weismer, G. (2011). An acoustic study of the relationships among neurologic disease, dysarthria type, and severity of dysarthria. *Journal of Speech, Language, and Hearing Research, 54*(2), 417–429. https://doi.org/10.1044/1092-4388(2010/10-0020)

Kim, Y-J., Weismer, G., Kent, R. D., & Duffy, J. R. (2009). Statistical models of F2 slope in relation to severity of dysarthria. *Folia Phoniatrica et Logopaedica, 61*(6), 329–335. https://doi.org/10.1159/000252849

King, J. M., Watson, M., & Lof, G. L. (2012). Practice patterns of speech-language pathologists assessing intelligibility of dysarthric speech. *Journal of Medical Speech-Language Pathology, 20*(1), 1–10.

Lam, J., & Tjaden, K. (2016). Clear speech variants: An acoustic study in Parkinson's disease. *Journal of Speech, Language, and Hearing Research, 59*(4), 631–646. https://doi.org/10.1044/2015_JSLHR-S-15-0216

Lansford, K. L., & Liss, J. M. (2014a). Vowel acoustics in dysarthria: Mapping to perception. *Journal of Speech, Language, and Hearing Research, 57*(1), 68–80. https://doi.org/10.1044/1092-4388(2013/12-0263)

Lansford, K. L., & Liss, J. M. (2014b). Vowel acoustics in dysarthria: Speech disorder diagnosis and classification. *Journal of Speech, Language, and Hearing Research, 57*(1), 57–67. https://doi.org/10.1044/1092-4388(2013/12-0262)

Lee, J., Littlejohn, M. A., & Simmons, Z. (2016). Acoustic and tongue kinematic vowel space in speakers with and without dysarthria. *International Journal of Speech-Language Pathology, 27*(3), 195–204.

Liss, J. M., Utianski, R., & Lansford, K. (2013). Crosslinguistic application of English-centric rhythm descriptors in motor speech disorders. *Folia Phoniatrica et Logopaedica, 65*(1), 3–19. https://doi.org/10.1159/000350030

Liu, H.-M., Tsao, F.-M., & Kuhl, P. K. (2005). The effect of reduced vowel working space on speech intelligibility in Mandarin-speaking young adults with cerebral palsy. *The Journal of the Acoustical Society of America, 117*(6), 3879–3889. https://doi.org/10.1121/1.1898623

Lüdecke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., & Makowski, D. (2021). performance: An R package for assessment, comparison and testing of statistical models. *The Journal of Open Source Software, 6*(60), 3139. https://doi.org/10.21105/joss.03139

McRae, P. A., Tjaden, K., & Schoonings, B. (2002). Acoustic and perceptual consequences of articulatory rate change in Parkinson disease. *Journal of Speech, Language, and Hearing Research, 45*(1), 35–50. https://doi.org/10.1044/1092-4388(2002/003)

Milenkovic, P. (2004). *TF32* [Computer program]. University of Wisconsin–Madison. https://ubeam.engr.wisc.edu/

Neel, A. T. (2008). Vowel space characteristics and vowel identification accuracy. *Journal of Speech, Language, and Hearing Research, 51*(3), 574–585. https://doi.org/10.1044/1092-4388(2008/041)

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance, 17*, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

Qualtrics. (2022). *Qualtrics.* https://www.qualtrics.com

R Core Team. (2022). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. https://www.R-project.org/

Rosen, K. M., Goozée, J. V., & Murdoch, B. E. (2008). Examining the effects of multiple sclerosis on speech production: Does phonetic structure matter? *Journal of Communication Disorders, 41*(1), 49–69. https://doi.org/10.1016/j.jcomdis.2007.03.009

Roussel, J.-R., Barber, C., Habel, K., Grasman, R., Gramacy, R. B., Mozharovskyi, P., & Sterratt, D. C. (2022). *geometry: Mesh generation and surface tessellation.* https://cran.r-project.org/web/packages/geometry/index.html

Rusz, J., Klempíř, J., Tykalová, T., Baborová, E., Čmejla, R., Růžička, E., & Roth, J. (2014). Characteristics and occurrence of speech impairment in Huntington's disease: Possible influence of antipsychotic medication. *Journal of Neural Transmission, 121*(12), 1529–1539. https://doi.org/10.1007/s00702-014-1229-8

Sandoval, S., Berisha, V., Utianski, R. L., Liss, J. M., & Spanias, A. (2013). Automatic assessment of vowel space area. *The Journal of the Acoustical Society of America, 134*(5), EL477–EL483. https://doi.org/10.1121/1.4826150

Sapir, S., Ramig, L. O., Spielman, J. L., & Fox, C. (2010). Formant centralization ratio: A proposal for a new acoustic measure of dysarthric speech. *Journal of Speech, Language, and Hearing Research, 53*(1), 114–125. https://doi.org/10.1044/1092-4388(2009/08-0184)

Skodda, S., Visser, W., & Schlegel, U. (2011). Vowel articulation in Parkinson's disease. *Journal of Voice, 25*(4), 467–472. https://doi.org/10.1016/j.jvoice.2010.01.009

Stipancic, K. L., Tjaden, K., & Wilding, G. (2016). Comparison of intelligibility measures for adults with Parkinson's disease, adults with multiple sclerosis, and healthy controls. *Journal of Speech, Language, and Hearing Research, 59*(2), 230–238. https://doi.org/10.1044/2015_JSLHR-S-15-0271

Story, B. H., & Bunton, K. (2017). Vowel space density as an indicator of speech performance. *The Journal of the Acoustical Society of America, 141*(5), EL458–EL464. https://doi.org/10.1121/1.4983342

Tjaden, K., Lam, J., & Wilding, G. (2013). Vowel acoustics in Parkinson's disease and multiple sclerosis: Comparison of clear, loud, and slow speaking conditions. *Journal of Speech, Language, and Hearing Research, 56*(5), 1485–1502. https://doi.org/10.1044/1092-4388(2013/12-0259)

Tjaden, K., Sussman, J., & Wilding, G. E. (2014). Impact of clear, loud, and slow speech on scaled intelligibility and speech severity in Parkinson's disease and multiple sclerosis. *Journal of Speech, Language, and Hearing Research, 57*(3), 779–792. https://doi.org/10.1044/2014_JSLHR-S-12-0372

Tjaden, K., & Wilding, G. E. (2004). Rate and loudness manipulations in dysarthria: Acoustic and perceptual findings. *Journal of Speech, Language, and Hearing Research, 47*(4), 766–783. https://doi.org/10.1044/1092-4388(2004/058)

Turner, G. S., Tjaden, K., & Weismer, G. (1995). The influence of speaking rate on vowel space and speech intelligibility for individuals with amyotrophic lateral sclerosis. *Journal of Speech and Hearing Research, 38*(5), 1001–1013. https://doi.org/10.1044/jshr.3805.1001

Weismer, G., Jeng, J.-Y., Laures, J. S., Kent, R. D., & Kent, J. F. (2001). Acoustic and intelligibility characteristics of sentence production in neurogenic speech disorders. *Folia Phoniatrica et Logopaedica, 53*(1), 1–18. https://doi.org/10.1159/000052649

Weismer, G., Yunusova, Y., & Bunton, K. (2012). Measures to evaluate the effects of DBS on speech production. *Journal of Neurolinguistics, 25*(2), 74–94. https://doi.org/10.1016/j.jneuroling.2011.08.006

Whitfield, J. A., Dromey, C., & Palmer, P. (2018). Examining acoustic and kinematic measures of articulatory working space: Effects of speech intensity. *Journal of Speech, Language, and Hearing Research, 61*(5), 1104–1117. https://doi.org/10.1044/2018_JSLHR-S-17-0388

Whitfield, J. A., & Goberman, A. M. (2014). Articulatory–acoustic vowel space: Application to clear speech in individuals with Parkinson's disease. *Journal of Communication Disorders, 51*, 19–28. https://doi.org/10.1016/j.jcomdis.2014.06.005

Whitfield, J. A., & Mehta, D. D. (2019). Examination of clear speech in Parkinson disease using measures of working vowel space. *Journal of Speech, Language, and Hearing Research, 62*(7), 2082–2098. https://doi.org/10.1044/2019_JSLHR-S-MSC18-18-0189

Wickham, H., & Wickham, M. H. (2016). *scales: Scale functions for visualization.* https://mran.microsoft.com/snapshot/2016-06-13/web/packages/scales/index.html

Zraick, R. I., Kempster, G. B., Connor, N. P., Thibeault, S., Klaben, B. K., Bursac, Z., Thrush, C. R., & Glaze, L. E. (2011). Establishing validity of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V). *American Journal of Speech-Language Pathology, 20*(1), 14–22. https://doi.org/10.1044/1058-0360(2010/09-0105)