JSLHR

Research Article

Assessing Vowel Centralization in Dysarthria: A Comparison of Methods

Annalise R. Fletcher,[a] Megan J. McAuliffe,[a] Kaitlin L. Lansford,[b] and Julie M. Liss[c]

Purpose: The strength of the relationship between vowel centralization measures and perceptual ratings of dysarthria severity has varied considerably across reports. This article evaluates methods of acoustic-perceptual analysis to determine whether procedural changes can strengthen the association between these measures.
Method: Sixty-one speakers (17 healthy individuals and 44 speakers with dysarthria) read a standard passage. To obtain acoustic data, 2 points of formant extraction (midpoint and articulatory point) and 2 frequency measures (Hz and Bark) were trialed. Both vowel space area and an adapted formant centralization ratio were calculated using first and second formants of speakers' corner vowels. Twenty-eight listeners rated speech samples using different prompts: one with a focus on intelligibility, the other on speech precision.
Results: Perceptually, listener ratings of speech precision provided the best index of acoustic change. Acoustically, the combined use of an articulatory-based formant extraction point, Bark frequency units, and the formant centralization ratio was most effective in explaining perceptual ratings. This combination of procedures resulted in an increase of 17% to 27% explained variance between measures.
Conclusions: The procedures researchers use to assess articulatory impairment can significantly alter the strength of relationship between acoustic and perceptual measures. Procedures that maximize this relationship are recommended.

Acoustic analysis of vowel sounds offers an objective assessment tool for measuring speech production in people with dysarthria. However, there are significant limitations in using acoustic metrics to infer information about listeners' perceptions of the disorder. Although studies have consistently reported an association between acoustic vowel centralization and perceptual measures, the strength of these relationships is highly variable (Lansford & Liss, 2014a). Linking measurements of the speech signal to perceptual outcomes is an important component of validating acoustic metrics for clinical use. Understanding causes of variation in the relationship between acoustic and perceptual data is a first step toward establishing stronger links between these variables.

Centralization of vowel formants has been associated with reduced intelligibility in both healthy speakers and those with motor speech disorders (e.g., Ferguson & Kewley-Port, 2007; Liu, Tsao, & Kuhl, 2005; Neel, 2008; Tjaden & Wilding, 2004). In the motor speech literature, the most common way of measuring vowel centralization is through the calculation of vowel space area (VSA)—using the first and second formants of a dialect's corner vowels. Static vowel formant values can be extracted across a range of word tokens, enabling measurements to be taken from a variety of speech stimuli. Unfortunately, VSA measurements have high interspeaker variability and have traditionally demonstrated variable success in distinguishing healthy and disordered speech (Sapir, Ramig, Spielman, & Fox, 2010). Indeed, VSA has been reported to account for both between 6% to 8% (Tjaden & Wilding, 2004) and 69% (H. Kim, Hasegawa-Johnson, & Perlman, 2011) of the variance in perceptual ratings of dysarthric speech (for a more detailed review, see Lansford & Liss, 2014a).

As Lansford and Liss (2014a) speculate, much of this inconsistency may be due to differences in the perceptual characteristics of participants' dysarthria from one study to another (e.g., there have been large differences in the

[a]Department of Communication Disorders and New Zealand Institute of Language, Brain & Behaviour, University of Canterbury, Christchurch
[b]School of Communication Science & Disorders, Florida State University, Tallahassee
[c]Department of Speech and Hearing Science, Arizona State University, Tempe
Correspondence to Annalise R. Fletcher: annalise.fletcher@canterbury.ac.nz

Editor: Joan Sussman
Associate Editor: Joan Sussman
Received October 9, 2015
Revision received April 11, 2016
Accepted July 15, 2016
DOI: 10.1044/2016_JSLHR-S-15-0355

Disclosure: *The authors have declared that no competing interests existed at the time of publication.*

severity of dysarthria exhibited by participants across studies). However, we hypothesize that this is not the only cause. Across studies, there are considerable differences in research methods used, and the contribution of these methods to results has thus far been overlooked. This article will compare procedures used to measure both vowel centralization and listeners' perceptions of dysarthric speech. The aim of the study is twofold. First, we will determine whether, and to what degree, changes in procedures affect the relationship between vowel centralization measurements and perceptual ratings. Second, to make recommendations for future studies, we will determine which set of procedures produces the strongest relationship between the acoustic and perceptual measurements. To accomplish this, we will evaluate the following techniques known to vary across studies: (a) the time point at which formant extraction occurs, (b) the method by which vowel centralization is calculated, and (c) the cues provided to listeners in their perceptual measurement of speech disorder.

Time Point of Formant Extraction

In the motor speech literature, formant measurements are almost universally taken from vowels' temporal midpoints. The rationale being that this provides a consistent measurement point that is as temporally removed from adjacent consonants as possible. However, it is well recognized that neighboring consonants can affect formant values across the entire vowel segment (Hillenbrand, Clark, & Nearey, 2001), and for this reason, the temporal midpoint may not necessarily provide the best representation of a vowel's steady-state formant frequency. In addition, Weismer and Berry (2003) demonstrated that the shape of formant movements can vary from speaker to speaker. This suggests that speakers might reach a vowel's steady-state target—or an approximation of this position—at different stages of the vowel's duration. If this is the case, the use of midpoint vowel measurements may obscure differences in formant movement between speakers.

To address this issue, more flexible measurement point criteria have been suggested (Fletcher, McAuliffe, Lansford, & Liss, 2015). Flexible criteria would enable us to extract an approximation of the vowel's steady-state target, irrespective of the time point that it is reached, for example, extracting formant values when they reach a particular stage of production, such as their maximum or minimum value. However, this approach has not been explored in the study of speech production in dysarthria. Thus, although we suspect that a flexible formant measurement point may be more successful in indexing speakers' articulatory impairment, there are currently no data to support this hypothesis.

Methods Used to Calculate Indices of Vowel Centralization

Acoustic metrics should index speech motor impairment while limiting the degree of interspeaker variation that is unrelated to speech disorder. However, regardless of the time point from which they are extracted, static vowel formants will always be affected by inherent differences in the size of speakers' vocal tracts. This variation obscures differences in vowel production that are due to changes in articulatory movement.

A number of methods aim to normalize anatomical and physiological differences between speakers' vowel formants. The intent is to minimize differences in formant measures that arise from the variables of age and sex (Clopper, 2009). However, in the study of motor speech disorders, many of these techniques can introduce problems. For example, normalizing the distances between speakers' vowels has the potential to remove information about the degree of articulatory movement made (i.e., as the speaker moves from the production of one vowel to another). In fact, some of the VSA differences observed between healthy male and female speakers may simply reflect sociolinguistic differences in articulatory movement—with women seeking to expand the acoustic distance between their vowels (Cox, 2006; Diehl, Lindblom, Hoemeke, & Fahey, 1996).

One method of vowel normalization—which reduces variance caused by the size of the vocal tract—is to transform the frequency scale used to measure formants (Clopper, 2009). Transformations of frequency measurements are classed as vowel-intrinsic methods of normalization and use only acoustic information contained within a single vowel to alter its formants. The aim of these methods, broadly speaking, is to model human vowel perception—not to eliminate physiological differences between speakers. Measuring formants in the Bark frequency scale has been shown to reduce the absolute difference between healthy male and female VSAs (as demonstrated by differences in two analyses of Hillenbrand, Getty, Clark, & Wheeler's, 1995, data set; see Neel, 2008; Sapir et al., 2010). However, it is unclear whether this also reduces interspeaker differences in articulatory impairment. H. Kim et al. (2011) found a strong relationship (i.e., $R^2 = .69$) between measurements of VSA in Bark and intelligibility scores for speakers with cerebral palsy. Although the study did not directly compare different units of measurement, the relationship between triangular VSA and intelligibility measurements was the strongest of all studies of dysarthric speech reviewed by Lansford and Liss (2014a), suggesting that there may be advantages to using Bark frequency units.

Measuring in Bark frequency units is not the only means by which normalization may be achieved. In a recent article, Sapir et al. (2010) suggested that using a ratio of each person's formant values would normalize interspeaker variance in the magnitude of formant values while preserving information about vowel centralization. They advocated use of the Formant Centralization Ratio (FCR), which weighs formants that are likely to increase as a result of vowel centralization against formants that are expected to lower (Sapir et al., 2010). This was supported by Lansford and Liss (2014a), who found that FCR produced a stronger correlation between vowel centralization and listeners' perceptions of dysarthric speech than traditional measures of VSA. Using the same formants, the FCR measure was

able to account for 15% more of the variance in speakers' intelligibility than a triangular VSA. Although these results were promising, data on this new measurement tool are lacking. It is not yet clear how the FCR compares with other vowel-intrinsic methods of vocal tract normalization (e.g., VSA measured in Bark) or, perhaps more importantly, whether the FCR is able to consistently index listeners' perception of dysarthria severity.

Perceptual Measurement of Speech Disorder

Acoustic metrics of vowel production are commonly indexed against some form of speech intelligibility measurement, for example, listener transcriptions of words and phrases (H. Kim et al., 2011; Lansford & Liss, 2014a; Liu et al., 2005) or scaled ratings of intelligibility (e.g., Turner, Tjaden, & Weismer, 1995; Weismer, Jeng, Laures, Kent, & Kent, 2001). Measuring dysarthria perceptually allows researchers to make inferences about the effects of the disorder on everyday communication. Yet, although orthographic transcription provides information regarding the proportion of words a listener has understood, its ability to detect and account for mild articulatory impairment can be limited (Sussman & Tjaden, 2012). That is, a listener may exhibit a perceptible dysarthria but achieve similar scores to healthy speakers on transcription intelligibility tests. Rating scales offer a useful alternative—allowing listeners to indicate that they detect speech impairment, even if they can still understand the words spoken. A number of studies have reported a relationship between scaled ratings of intelligibility and the degree of vowel centralization evidenced by individuals with dysarthria (Y. Kim, Kent, & Weismer, 2011; McRae, Tjaden, & Schoonings, 2002; Tjaden & Wilding, 2004; Turner et al., 1995; Weismer et al., 2001). However, the strength of these relationships remains highly variable, with VSA accounting for anywhere between 6% and 46% of the variance in listener ratings. In addition to the different procedures used to measure vowel centralization, the instructions provided to listeners as they complete a rating task might contribute to this variation.

When rating scales are used to measure speech impairment in dysarthria, listeners are usually asked to rate intelligibility or "how easy" the speaker is to understand (Y. Kim et al., 2011; Tjaden & Wilding, 2004; Turner et al., 1995; Weismer et al., 2001). However, it is possible that this approach to intelligibility rating might—to some degree—be prone to the same issues as transcription-based intelligibility scores. That is, even when listeners detect mild articulatory impairment, they may still rate a speaker as very easy to understand. To combat this issue, generalized ratings of speech severity have been proposed, with the idea that these may index speech impairment not adequately captured through measures of word or sentence intelligibility (Sussman & Tjaden, 2012). Indeed, Sussman and Tjaden (2012) found that scaled estimates of speech severity were able to distinguish speakers with mild dysarthria more successfully than transcription-based intelligibility scores. Although the study suggested that the instructions we give listeners are important in measuring dysarthria, it did not directly compare different listener prompts (i.e., prompts to rate "intelligibility" vs. prompts to rate "speech severity"). Hence, it is not clear whether the instruction to rate "speech severity"—as opposed to "intelligibility"—made any difference to the sensitivity of their rating scale.

There are limited data to evaluate how listener instructions affect the measurement of dysarthric speech. Weismer et al. (2001) compared ratings of "intelligibility" with "speech severity" and found little difference in the amount that each rating predicted acoustic changes in VSA. However, in rating speech intelligibility, listeners in the Weismer et al. (2001) study were told to focus on articulatory precision. It is possible that by focusing on articulatory precision, listeners produced ratings that were more sensitive to mild dysarthria. In contrast, instructions to rate speech severity asked the listener to focus on all aspects of possible speech disorders including parameters of nasality, prosody, vocal quality, and respiration. Although these parameters are likely to be affected by the presence of dysarthria, they do not directly influence vowel centralization. For this reason, to best index changes in acoustic vowel production, it may be beneficial to have listeners rate a speaker's speech precision irrespective of other speech subsystem impairment.

In summary, there are a number of methodological factors that might affect the relationship between vowel centralization and listeners' perceptions of dysarthric speech. However, it is unclear to what degree these factors are capable of changing this relationship and therefore contributing to the variable results reported in previous studies. The current study will evaluate a variety of methods of acoustic and perceptual analysis to determine what effect measurement differences have on the relationship between vowel centralization and listeners' perceptions of dysarthric speech. In doing so, this study aims to determine which measures produce the strongest relationship between these variables and therefore provide the clearest acoustic index of dysarthria severity. Specifically, this investigation will compare the use of different (a) formant extraction time points (midpoint and a flexible measurement point), (b) methods of vocal tract normalization (Hertz and Bark), and (c) listener rating cues (speech intelligibility and speech precision). The results address several questions: (a) Do methodological changes produce significantly different vowel dispersion values and perceptual rating outcomes? (b) Are the resultant measurements able to distinguish individuals with dysarthria from healthy older speakers? (c) Do these methodological changes strengthen the relationship between the acoustic and perceptual measures?

Method

Speakers

Sixty-one speakers of New Zealand English (NZE; 42 men and 19 women), aged between 43 and 89 years, participated in this study. Of these speakers, 44 were diagnosed with dysarthria. The dysarthria varied in severity, with speakers classed as exhibiting mild (n = 16), mild-moderate

(n = 9), moderate (n = 8), moderate-severe (n = 4), and severe (n = 7) dysarthria. Perceptual classification of severity was rated by three experienced speech-language pathologists via a consensus rating procedure, on the basis of speakers' recordings of the Grandfather Passage (see Appendix A). Biographical details are supplied in Table 1. The remaining 17 speakers, who reported no history of neurological impairment or speech and language disorders, acted as healthy controls. The group diagnosed with dysarthria had a mean age of 65 years, whereas the control group had a mean age of 66 years.

Speech Stimuli

Each speaker attended a single recording session. Recordings took place in a quiet room, with an investigator present. Participants were asked to read the Grandfather Passage in their normal speaking voice after familiarizing themselves with the passage. Two participants with dysarthria required assistance reading the passage. In these instances, the first author read full sentences from the passage, with the speaker repeating the sentences immediately afterward. For 58 participants, digital audio recordings were made via an Audix HT2 headset condenser microphone, positioned approximately 5 cm from the mouth. Digital audio recordings of these speakers were made at 48 kHz with 16 bits of quantization. The remaining three participants were female control speakers who were recorded as part of an earlier study. These participants were recorded using a Zoom H4n recorder placed on the table in front of them (at an approximate distance of 30 cm). Their audio

Table 1. Demographic information for speakers with dysarthria.

Participant number	Sex	Age	Medical etiology	Severity of disorder
1	F	48	Traumatic brain injury	Mild-moderate
2	M	60	Traumatic brain injury	Moderate
3	M	55	Traumatic brain injury	Mild-moderate
4	F	67	Progressive supranuclear palsy	Mild
5	F	68	Freidreich's ataxia	Mild
6	F	70	Parkinson's disease	Mild-moderate
7	M	75	Parkinson's disease	Moderate
8	F	79	Parkinson's disease	Mild
9	M	56	Cerebellar ataxia	Mild
10	F	45	Wilson's disease	Mild
11	M	53	Undetermined neurological disease	Moderate
12	M	55	Undetermined neurological disease	Moderate
13	M	58	Brainstem stroke	Moderate
14	M	76	Parkinson's disease	Mild
15	M	67	Parkinson's disease	Mild-moderate
16	M	77	Parkinson's disease	Mild
17	M	67	Parkinson's disease	Mild
18	M	79	Parkinson's disease	Moderate
19	M	71	Parkinson's disease	Moderate
20	M	71	Parkinson's disease	Mild-moderate
21	F	83	Parkinson's disease	Mild
22	M	68	Parkinson's disease	Mild
23	F	73	Parkinson's disease	Mild-moderate
24	M	89	Parkinson's disease	Mild
25	M	58	Parkinson's disease	Mild
26	M	81	Parkinson's disease	Moderate-severe
27	M	73	Parkinson's disease	Mild
28	M	79	Parkinson's disease	Mild
29	M	77	Parkinson's disease	Moderate-severe
30	M	69	Parkinson's disease	Moderate
31	M	69	Parkinson's disease	Mild
32	M	65	Parkinson's disease	Mild-moderate
33	M	68	Parkinson's disease	Mild
34	M	47	Traumatic brain injury	Severe
35	M	64	Spinocerebellar ataxia	Severe
36	F	69	Cerebral palsy	Severe
37	F	60	Multiple sclerosis	Moderate-severe
38	M	55	Huntington's disease	Severe
39	F	53	Multiple sclerosis	Mild-moderate
40	F	47	Huntington's disease	Moderate-severe
41	M	43	Hydrocephalus	Severe
42	M	60	Cerebral palsy	Severe
43	M	72	Stroke	Severe
44	F	46	Brain tumor	Mild-moderate

Note. F = female; M = male.

recordings were made at 22.05 kHz with 16 bits of quantization. As part of the formant extraction procedure, all sound files were later resampled to a lower frequency as per the Burg linear predictive coding algorithm described in the next section.

Extraction of Acoustic Data

Segmentation of the Data Set

The recordings were transcribed, automatically segmented to the phoneme level, and labeled in Praat (Boersma & Weenink, 2012) using the Hidden Markov Model Toolkit (Young et al., 2002). Phoneme segments were labeled in Praat on the basis of the origins of NZE miner orthographic-phonemic dictionary (Fromont & Hay, 2008), constructed from the Celex lexical database (Baayen, Piepenbrock, & Gulikers, 1996) and additional hand-labeled entries. The accuracy of all phoneme boundaries was checked by a team of four trained analyzers who visually examined the waveform and wide-band spectrogram and listened for auditory cues. The primary indicators for the onset and offset of vowels were changes to formant structures, voicing, and waveform amplitude. Vowel onset boundaries were identified at the start of the pitch period, coinciding with the onset of regular formant structure. Vowel offset boundaries were distinguished by changes in formant structure at the end of the pitch period, where there was a corresponding drop in waveform amplitude. The amplitude, shape, and lack of frication of successive pitch periods were also used to determine boundaries. Because the Hidden Markov Model Toolkit segmentation was completed at the phoneme level, if the person checking phoneme boundaries was uncertain in discriminating boundaries for consecutive phonemes, the boundary derived from automatic segmentation was kept in place.

Extraction of Formant Values

Three tokens of the NZE START [ɐː], FLEECE [iː], and THOUGHT [oː] vowels were selected from the passage for the calculation of acoustic metrics. These tokens tend to elicit the most extreme front [iː], open [ɐː], and back [oː] vowel positions in NZE. The [ɐː] vowel was extracted from two occurrences of the word *grandfather*[1] and one occurrence of the word *answers*.[2] The [iː] vowel was extracted from two occurrences of the word *each* and one occurrence of the word *three*. The [oː] vowel was extracted from one occurrence of the words *organ, short,* and *more*. Because of reading errors, speakers occasionally missed one of the selected tokens. In this instance, the remaining two tokens were used. In instances of dysfluency, where speakers repeated certain word tokens, the average formant value across word repetitions was used. The formant tracks of the first five formant frequencies were obtained via Praat using the Burg linear predictive coding algorithm, with a Gaussian window length of 25 ms, a time step of 6.25 ms between the centers of consecutive windows, a maximum formant value of 5.5 kHz for women and 5 kHz for men, and a preemphasis from 50 Hz (Boersma & Weenink, 2012). Formant 1 (F1) and Formant 2 (F2) measurements were extracted from two measurement points in each vowel. Criteria for the formant measurement points are outlined below. Each set of vowel formants was measured in Hz and also transformed into the Bark frequency scale (Traunmüller, 1990).

Midpoint Formant Values

Midpoint values were automatically extracted using a custom Praat script. All formant tracks were also visually checked. If the midpoint values selected by the script did not accurately represent the formant that was being measured (i.e., the formant track was not centered on the correct formant band), the measurement point was adjusted by hand.

Articulatory Point Formant Values

The articulatory point criteria were designed with the aim of extracting values at the time at which there was the least movement in the formant tracks—for the best approximation of the vowels' steady-state target. For the front [iː] vowel, this point was set at peak F2 frequency; for the open [ɐː], formants were extracted when F1 was at its maximum; and for back [oː] vowel, when the lowest value of F2 was reached. Articulatory point formant values were all automatically extracted using a custom Praat script. All vowel measurement points were then visually checked. As described above, if the values selected by the script did not accurately represent the formant that was being measured, the measurement point was adjusted by hand. An example of how the midpoint and articulatory extraction points might differ is shown visually in Figure 1.

Description of the Acoustic Metrics

Vowel centralization was calculated with two metrics —VSA and FCR. The measures used are described below.

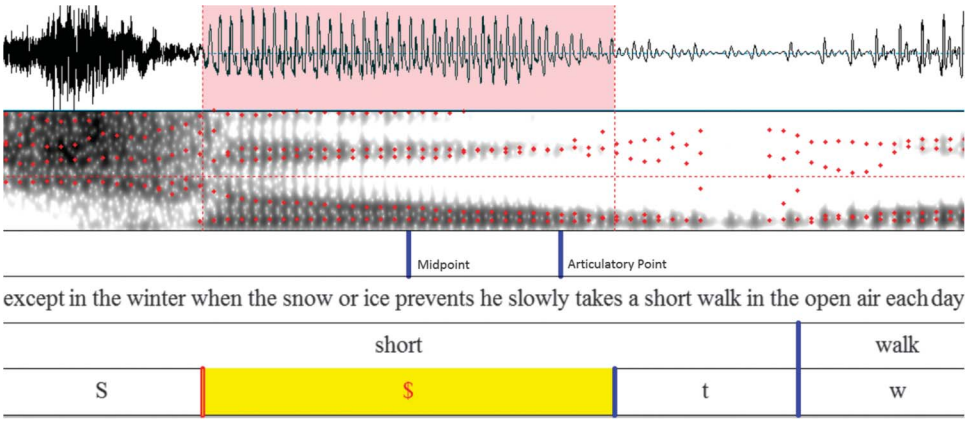
VSA

VSA was calculated using F1 and F2 of the [ɐː], [iː], and [oː] vowels. Given the NZE dialect, a measure of triangular VSA (using the THOUGHT vowel as opposed to GOOSE) provides a more accurate representation of vowel dispersion than quadrilateral VSA (Maclagan, 2009). F1 and F2 values for the three [ɐː], [iː], and [oː] word tokens were averaged for each speaker. Triangular VSA was constructed by plotting these values as coordinates in an F1/F2 plane and calculating the resulting triangular area using the formula $Hz^2 = 0.5 \times ABS[F1[iː] \times (F2[ɐː] - F2[oː]) + F1[oː] \times (F2[iː] - F2[ɐː]) + F1[ɐː] \times (F2[oː] - F2[iː])]$, where ABS = absolute value, F1[iː] = first formant frequency of the [iː] vowel, and so on.

[1]The primary stress in *grandfather* usually occurs on the first syllable; however, there was always adequate stress on the second syllable to produce a distinctive [ɐː] token.

[2]In NZE, *answers* always contains a START vowel rather than a TRAP vowel.

Figure 1. Example of the two extraction points within a speaker's [o:] vowel.



Formant Centralization Ratio

Given the dispersion of vowels in NZE, the FCR metric was adapted from Sapir et al. (2010).[3] It was calculated with the same average F1 and F2 values for each speaker as detailed above, again using the THOUGHT vowel as opposed to GOOSE. Therefore, FCR was realized as (F2[oː] + F2[ɐː] + F1[iː] + F1 [ɐː]) ÷ (F2[iː] + F1[oː]).

The procedures described resulted in eight different formant centralization measurements for each speaker, outlined in Table 2.

Reliability of Acoustic Measures

To determine inter- and intrarater reliability of the measures, 10% of text grids were manually reexamined for reliability. Phoneme boundaries were manually rechecked, and scripts to obtain vowel formant values were readministered. The newly generated vowel formants values were visually checked in the same manner as the original values. In the case of midpoint formant values, the reanalysis found F1 intrarater reliability scores averaged within 12 Hz of original values, and F2 scores were within 22 Hz. The average interrater difference was 26 Hz for F1 values and 46 Hz for F2. The reanalysis of the articulatory points found F1 intrarater reliability scores within 16 Hz of original values and F2 scores within 23 Hz. Average interrater differences were 35 Hz for F1 values and 29 Hz for F2.

Perceptual Task

To gather perceptual ratings, two different prompts were used, with two listener groups.

[3]The formula provided by Sapir et al. (2010) is given as (F2/u/ + F2/ɑ/ + F1/i/ + F1/u/) / (F2/i/ + F1/ɑ/). In NZE, the vowel in THOUGHT is produced much further back than the vowel in GOOSE. For this reason, its inclusion better represents the overall vowel dispersion of our speakers.

Listeners

Listeners consisted of two randomly assigned groups of 14 adults (aged 18 to 47 years). The listeners were native speakers of NZE, who were unfamiliar with dysarthric speech. All listeners passed a pure-tone hearing screening at 20 dB hearing level for 500, 1,000, 2,000, and 4,000 Hz in both ears.

Listening Stimuli

Because of the large amount of speech data collected in this study, only a small portion of the reading passage was used to gather perceptual ratings. The phrase "he slowly takes a short walk in the open air each day" was selected for this purpose. Across the speaker group, this phrase was free from reading errors. For all recordings, the average intensity of the phrase was scaled to 70 dB SPL to provide a similar perceived loudness.

Procedure

The two listener groups each completed one listening task. All listeners completed the rating task in one session. The two listening tasks were programmed in E-prime,

Table 2. Combinations of acoustic metrics.

Formant measurement point	Unit of measurement	Vowel centralization metric
Temporal midpoint	Hz	VSA
Temporal midpoint	Hz	FCR
Temporal midpoint	Bark	VSA
Temporal midpoint	Bark	FCR
Articulatory target	Hz	VSA
Articulatory target	Hz	FCR
Articulatory target	Bark	VSA
Articulatory target	Bark	FCR

Note. VSA = vowel space area; FCR = formant centralization ratio.

and speech stimuli were played through Panasonic RP-HT 161 stereo headphones. In Group 1, listeners were asked to rate how easy the speaker was to understand, whereas in Group 2, listeners were asked to rate the speakers' speech precision. Exact participant instructions are provided in Appendix B.

Although the two listener groups were given different task prompts, all other rating procedures were identical. Before beginning the experiment, listeners completed a short practice task to become familiar with the rating procedure and provide the opportunity to adjust the volume of the computer to a comfortable level. In the practice task, listeners were exposed to three recordings. These included a speaker with severe dysarthria, a speaker with mild-moderate dysarthria, and one healthy older speaker. These speakers were not included in the main experiment.

The main experiment consisted of 61 phrases—one from each speaker listed in Table 1. The phrases were randomly presented twice, giving a total of 122 trials for every listener. In every trial, the listeners were presented with a prompt to either rate "the speaker's speech precision" or "how easy is the speaker to understand?" Listeners pressed a button to hear the recording play and clicked on a visual analog scale (VAS) to place a copy of the button onto the scale. For listeners in Group 1, the scale ranged from "easy" at one end to "difficult" at the other. For the second group, the scale ranged from "precise" to "imprecise." Listeners were able to adjust their rating as often as they wished before selecting to move to the next trial.

The raw output of these judgments was an integer between 0 and 100 for each stimulus phrase. For each listener, we calculated the average and standard deviation of all the ratings provided. This information was used to compute a z score for every speaker that was rated by the listener. For example, a numeric rating given by Listener A would be converted in the following manner:

$$\frac{\text{rating of speaker by Listener A} - \text{average rating given by Listener A}}{\text{standard deviation of Listener A's ratings}}$$

This z score procedure ensured that listeners who tended to give speakers higher ratings (while placing bigger spaces between different speaker ratings on the VAS) would not have undue influence on the overall rating averages. After applying this z score procedure, the scores of all listeners were averaged to determine two final ratings for that speaker—one of speech intelligibility and another for speech precision.

Reliability of the Perceptual Task

To assess inter- and intrarater reliability, Pearson's product–moment correlations (across ratings of the same speech samples) were calculated on the basis of listeners' raw ratings (i.e., scores between 0 and 100). For intelligibility ratings, the average intrarater correlation between the first and second presentation of the phrases ranged from .70 to .95, with a mean of .88. For ratings of speech precision, the intrarater correlations were between .86 and .96, with a mean of .90. To assess interrater reliability, intraclass correlations (ICCs) were calculated (as described in Sheard, Adams, & Davis, 1991). The obtained ICC (2,1) coefficients were .677 for intelligibility ratings and .835 for speech precision ratings.

Results

The results of this study are discussed in three parts and address (a) whether measurements of vowel dispersion and perceptual ratings were affected by methodological changes, (b) whether these measurements were able to distinguish individuals with dysarthria from healthy older speakers, and (c) whether methodological changes strengthened the relationship between the acoustic and perceptual measures.

Effect of Methodological Changes on Vowel Dispersion and Perceptual Ratings

Variation in Method of Formant Extraction

The use of midpoint and flexible extraction methods resulted in statistically significant differences in the size of speakers' VSAs. In speakers with dysarthria, VSA calculated using F1 and F2 from the average midpoint was significantly smaller (M = 147,315 Hz^2, SD = 74,337 Hz^2) than VSA calculated using F1 and F2 extracted using our articulatory point measure (M = 207,575 Hz^2, SD = 86,283 Hz^2, t[43] = 11.2, $p < .001$). This was also the case for control speakers, with the midpoint formants producing significantly smaller VSA values (M = 217,220 Hz^2, SD = 81,373 Hz^2) than those extracted from the articulatory point (M = 293,539 Hz^2, SD = 106,344 Hz^2, t[16] = 7.7, $p < .001$). However, despite these differences, the two measures were highly correlated, r(59) = .93, $p < .001$. The same general pattern was true of the FCR measures, with smaller mean FCR values generated from the articulatory point formant values (see Table 3). Again, the correlation between the two FCR measures (taken in Hz) was high, r(59) = .94, $p < .001$.

Variation in Unit of Measurement and Vowel Centralization Metric

Table 3 provides mean FCR and VSA values of male and female speakers calculated using Hertz and Bark and across the two measurement points. The results indicate that, as expected, formant vales for men and women are more similar when measured in Bark. These data suggest that differences caused by the size of the vocal tract are indeed reduced when the Bark scale is used. Table 3 also demonstrates that the mean difference between men and women is reduced when vowel centralization is measured using the FCR as opposed to VSA. Together, the combined use of the Bark scale and the FCR eliminated any significant differences in vowel centralization measurements between male and female speaker groups—both when midpoint, t(59) = 1.05, $p > .05$, and articulatory point formant values were used, t(59) = 0.57, $p > .05$.

Table 3. Measurement differences between men and women.

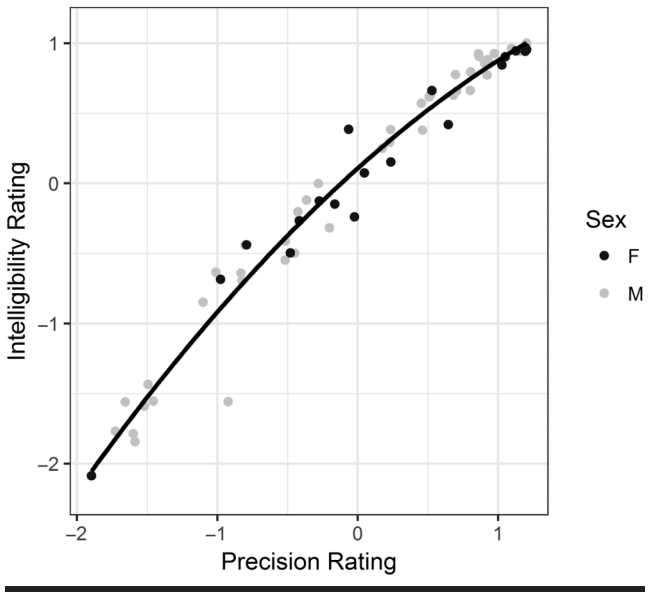
	VSA, Hz^2	VSA, Bark^2	FCR, Hz	FCR, Bark
Articulatory target measurement				
Male	195,295 (80,218)	8.493 (3.25)	1.048 (0.12)	1.265 (0.11)
Female	311,635 (91,527)	11.192 (2.89)	0.987 (0.08)	1.249 (0.07)
Temporal midpoint measurement				
Male	137,331 (65,552)	5.974 (2.60)	1.132 (0.12)	1.348 (0.11)
Female	231,932 (78,255)	8.394 (2.60)	1.061 (0.10)	1.319 (0.08)

Note. VSA = vowel space area; FCR = formant centralization ratio.

Variation in Instructions Provided to Listeners

The relationship between the two perceptual measurements—intelligibility and speech precision—is shown in Figure 2. Overall, listeners' perceptions, as measured by the two different rating instructions, were highly correlated, $r(59) = .98$, $p < .001$. Although closely related, the data points in Figure 2 appeared to have a curvilinear relationship. This observation was confirmed by comparing a simple linear regression against a second-degree polynomial model of the two variables. A comparison of the models revealed that the curvilinear, polynomial model accounted for significantly more variance in the data, $F(1, 58) = 21.03$, $p < .001$. The existence of a curvilinear relationship indicates that there were differences in the way the two sets of instructions provided to listeners indexed mild, moderate, and severe dysarthria. For example, speakers with a mild dysarthria tended to exhibit higher z scores (i.e., scores that were further away from the mean) for speech precision than for intelligibility. This was not the case for speakers with low ratings. This suggested that ratings of intelligibility and speech precision were distributed differently, with ratings of speech precision producing a larger range of scores for speakers above the mean (i.e., those with less impairment).

Figure 2. Relationship between listeners' ratings of intelligibility and speech precision. M = male; F = female.



Differences Between Speakers With and Without Dysarthria Across Acoustic and Perceptual Measures

Table 4 highlights differences in the perceptual and acoustic measurements between speakers with dysarthria and healthy controls. The perceptual measurements were combined for male and female speakers after determining that there were no significant differences between the sexes for ratings of intelligibility, $t(59) = 0.85$, $p > .05$, or speech precision, $t(59) = 0.95$, $p > .05$. Perceptual ratings of speech precision produced a greater mean difference between speakers with and without dysarthria than ratings of intelligibility (after both measures had been z scored).

The acoustic measures were compared separately in groups of male and female speakers. All measurements produced statistically significant differences between the speakers with and without dysarthria (at $p < .05$). However, it was apparent that some measures were able to separate the two groups more clearly than others (i.e., there was less overlap in the distribution of measurements across the two groups, as indicated by higher t values). First, formants taken with a flexible extraction point consistently produced higher t values in comparisons between the speakers with and without dysarthria. Measuring formant values in Bark units also produced consistently higher t values. In contrast, the FCR did not perform consistently better than measures of VSA in distinguishing speakers with dysarthria—as measures of VSA in Bark^2 produced particularly high t values in these group comparisons.

Changes to the Strength of the Relationship Between Acoustic and Perceptual Measures With Variation in Methodology

A series of Pearson correlation analyses were performed to evaluate whether the relationship between the vowel centralization and perceptual ratings of dysarthria varied with methodological changes. The results are summarized in Table 5. The data were analyzed separately by sex because this factor accounted for significant variation

Table 4. Differences in perceptual ratings and vowel dispersion metrics in participants with and without dysarthria.

Measurement	Average in speakers with dysarthria	Average in control speakers	*t*	*p*
Rating of speech precision	–0.382 (0.805)	0.987 (0.198)	6.895	<.001
Rating of intelligibility	–0.339 (0.834)	0.877 (0.110)	5.960	<.001
Male speakers				
VSA in $Bark^2$ using a flexible formant extraction point	7.635 (3.034)	10.912 (2.635)	3.178	.003
FCR using formants measured in Bark from a flexible extraction point	1.293 (0.113)	1.186 (0.066)	2.967	.005
VSA in $Bark^2$ using formants from the temporal midpoint	5.341 (2.479)	7.755 (2.156)	2.864	.007
FCR using formants measured in Hz from a flexible extraction point	1.077 (0.124)	0.968 (0.070)	2.746	.009
FCR using formants measured in Bark from the temporal midpoint	1.372 (0.111)	1.280 (0.064)	2.594	.01
VSA in Hz^2 using a flexible formant extraction point	178,294 (77,630)	243,206 (69,879)	2.441	.02
FCR using formants measured in Hz from the temporal midpoint	1.157 (0.128)	1.061 (0.078)	2.325	.03
VSA in Hz^2 using formants from the vowels' temporal midpoint	124,062 (64460)	174,724 (55396)	2.317	.03
Female speakers				
FCR using formants measured in Bark from a flexible extraction point	1.279 (0.060)	1.185 (0.028)	3.615	.002
VSA in $Bark^2$ using a flexible formant extraction point	9.958 (2.088)	13.866 (2.641)	3.496	.003
VSA in $Bark^2$ using formants from the temporal midpoint	7.287 (2.187)	10.794 (1.682)	3.464	.003
FCR using formants measured in Hz from a flexible extraction point	1.018 (0.071)	0.920 (0.031)	3.224	.005
FCR using formants measured in Bark from the temporal midpoint	1.351 (0.080)	1.248 (0.029)	3.024	.008
VSA in Hz^2 using a flexible formant extraction point	277,397 (64,148)	385,816 (103,108)	2.829	.01
VSA in Hz^2 using formants from the vowels' temporal midpoint	202,764 (68454)	295,128 (61,605)	2.814	.01
FCR using formants measured in Hz from the temporal midpoint	1.097 (0.104)	0.984 (0.029)	2.571	.02

Note. Standard deviations across groups are shown in parentheses. All *t* values were derived from two-sample, independent *t* tests. Equal variance between groups was assumed after applying Levene's test. Absolute *p* values are reported, with no corrections made for multiple comparisons. VSA = vowel space area; FCR = formant centralization ratio.

in speakers' acoustic measurements after controlling for perceptual ratings. This was not the case for other biographical factors, such as dysarthria etiology.[4] Table 5 shows that in both male and female speakers, there were common methodological approaches that improved the association between perceptual and acoustic measurements.

In combination, changes to the formant extraction point, unit of measurement, metric of vowel centralization, and listener instructions resulted in 17% more variance being accounted for in men (i.e., an increase from 17% to 34% when all four changes were made) and 27% more variance accounted for in women (an increase from 49% to 76%). Overall, the strongest relationship between acoustic and perceptual measures—in both male and female speakers—was achieved by using a flexible formant extraction point, Bark units, and the FCR metric, in combination with listener ratings of speech precision. Figure 3 plots the strongest and weakest relationships found between the acoustic and perceptual measures in male and female speakers.

Discussion

One of the main aims of this study was to determine which combination of procedures would result in the strongest relationship between measurements of vowel centralization and listeners' perceptions of dysarthria. Previous literature has reported considerable variability in the correlations between these measurements, and it is unclear to what degree different procedures might be contributing to this inconsistency. It was hypothesized that there were several methodological changes that might affect the relationship between our acoustic and perceptual measurements.

[4]In assessing this, we ran a series of stepwise multiple regressions to model each metric of vowel centralization. Model fitting proceeded in a forward stepwise iterative manner, seeking to produce a model containing only significant effects. Speakers' sex and their rating of speech precision produced the best models of their VSA values (regardless of which units and formant extraction point were used). In contrast, speech precision ratings were the only significant predictor of FCR values. Information about speakers' intelligibility, presence/absence of dysarthria, and dysarthria etiology provided no additional statistically significant information about their VSA or FCR measures (i.e., $p > .05$ for all additional variables added to the models).

Table 5. Relationships between acoustic vowel metrics and perceptual measures.

Vowel space metric	Speech precision rating		Intelligibility rating	
	Correlation coefficient	Explained variance (%)	Correlation coefficient	Explained variance (%)
Women				
FCR using formants measured in Bark from a flexible extraction point	–.873	76	–.839	70
FCR using formants measured in Bark from the temporal midpoint	–.854	73	–.810	66
FCR using formants measured in Hz from a flexible extraction point	–.852	73	–.836	70
VSA in $Bark^2$ using formants from the temporal midpoint	.832	69	.774	60
VSA in $Bark^2$ using a flexible formant extraction point	.808	65	.750	56
FCR using formants measured in Hz from the temporal midpoint	–.798	64	–.764	58
VSA in Hz^2 using formants from the vowels' temporal midpoint	.751	56	.710	50
VSA in Hz^2 using a flexible formant extraction point	.739	55	.698	49
Men				
FCR using formants measured in Bark from a flexible extraction point	–.584	34	–.565	32
FCR using formants measured in Hz from a flexible extraction point	–.581	34	–.568	32
VSA in $Bark^2$ using a flexible formant extraction point	.576	33	.556	31
FCR using formants measured in Bark from the temporal midpoint	–.550	30	–.523	27
FCR using formants measured in Hz from the temporal midpoint	–.548	30	–.531	28
VSA in Hz^2 using a flexible formant measurement point	.516	27	.500	25
VSA in $Bark^2$ using formants from the temporal midpoint	.505	26	.475	23
VSA in Hz^2 using formants from the temporal midpoint	.435	19	.407	17

Note. All correlations have p values less than .01. VSA = vowel space area; FCR = formant centralization ratio.

Method of Formant Extraction

Two different formant extraction points were explored: a static temporal midpoint and a flexible articulatory point. It was hypothesized that the articulatory point might better index speakers' articulatory impairment by capturing a larger degree of vocal tract movement between vowels. Consistent with previous work, VSA values were considerably larger when formants were extracted from the articulatory point (Fletcher et al., 2015). Overall, these results provide further support for the hypothesis that extracting F1 and F2 values at a flexible articulatory point produces more acoustically distinct formant values.

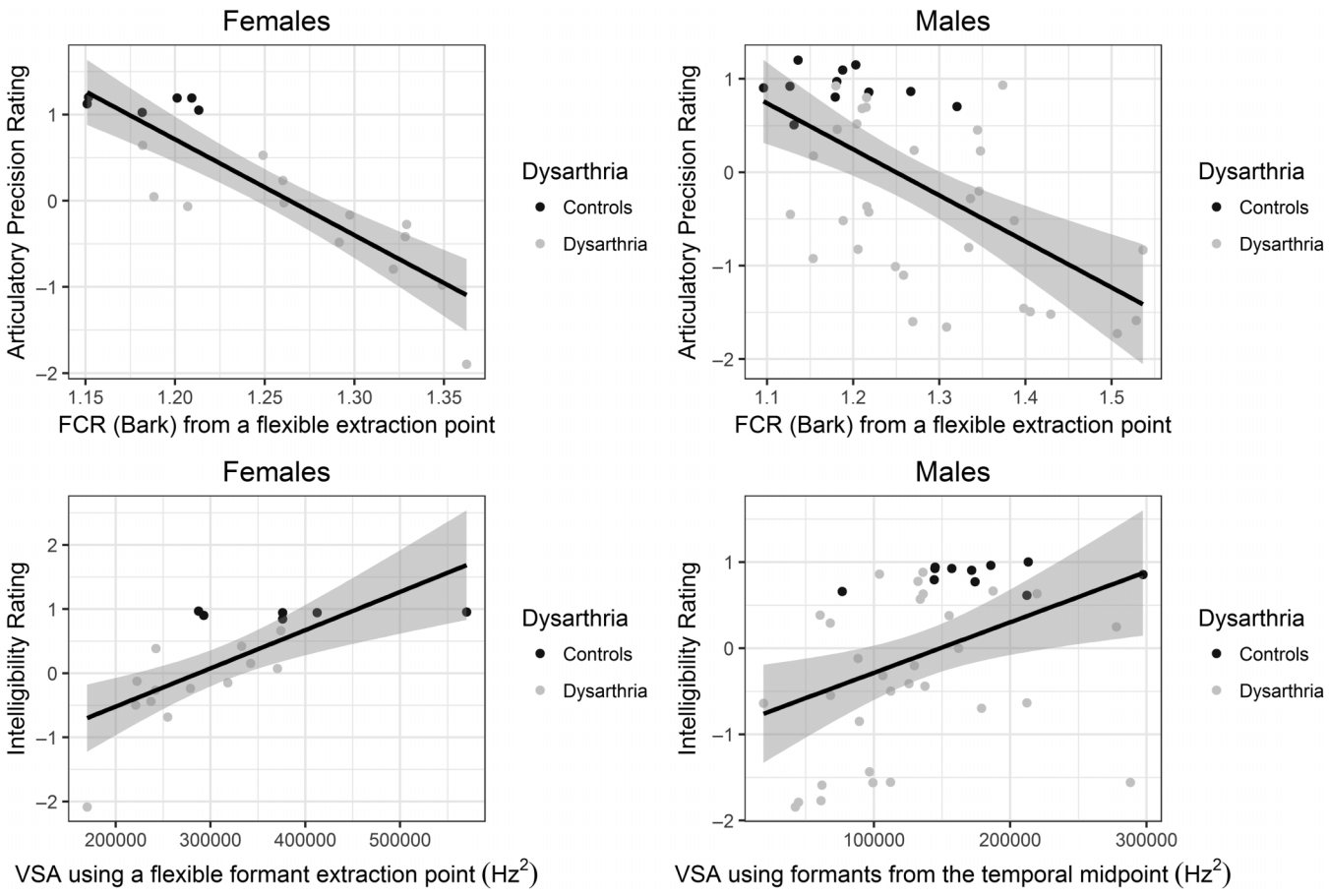
In extracting formants from the articulatory point, we aimed to determine whether this method would strengthen the relationship between vowel centralization metrics and perceptual ratings. The results indicated that formant measurements taken from the articulatory point tended to explain more of the variation in speakers' perceptions of dysarthria. The articulatory point formants were used in four vowel centralization metrics—applied to both male and female data. On average, they resulted in an increase of 6% in the variance accounted for by perceptual ratings in the male data but only 3% in the female data.

The reason for the decreased performance in the female data appears to be due to a single speaker. Figure 3 demonstrated that the relationship between the articulatory point VSA and listeners' perceptions of dysarthria was skewed by one high VSA value in the female subset. This outlier meant women's midpoint VSAs were more closely related to perceptual measures than the corresponding articulatory point VSAs. When the FCR was applied, the outlier was no longer apparent. These data suggest that articulatory point values are generally able to capture more information about speakers' perceived severity—but they may also be more sensitive to changes in vocal tract size. Hence, to achieve a strong relationship with perceptual ratings of dysarthria, it may be advisable to use articulatory point criteria in conjunction with the FCR procedure recommended by Sapir et al. (2010), to help normalize differences in vocal tract size.

Unit of Measurement and Vowel Centralization Metric

Outside of the motor speech disorder literature, it is common to convert formant data to Bark units before calculating VSA, in order to provide an appropriate auditory scaling of frequency (Ferguson & Kewley-Port, 2007; Neel, 2008). It was hypothesized that measuring vowel centralization using Bark units would increase the relationship between acoustic and perceptual measures by reducing the effect that differences in the size of the vocal tract had on speakers' VSAs. There is evidence that the use of Bark units did reduce

Figure 3. A comparison of the strongest and weakest relationships between acoustic vowel metrics and perceptual measures, plotted by sex. Shaded areas indicate 95% confidence interval of regression estimates.



these anatomical differences. In the current study, transformation of F1 and F2 to Bark units resulted in a reduction in the difference in VSA between male and female speakers (i.e., the measurements became less than one standard deviation apart). As expected, this also occurred when FCR was used in place of VSA. However, it was only when Bark units were applied to the FCR that the differences between the sexes became insignificant. This finding indicates that Bark units have the potential to reduce interspeaker variations over and above what the FCR alone is able to accomplish and that the use of the FCR does not necessarily render Bark units redundant.

Within our data set, the combined use of Bark units and the FCR provided a scale of values that could be interpreted together, regardless of a person's sex. The ability to plot and interpret these data as one group enhances sample size and increases the power to detect a relationship between our acoustic and perceptual measures. It is interesting that, in isolation, the application of the FCR was not able to completely eliminate group differences between men and women—in contrast to findings reported by Sapir et al. (2010) and Lansford and Liss (2014a). Differences may have persisted because of sociophonetic differences between the sexes (Cox, 2006; Diehl et al., 1996) or, simply, because complex differences in vocal tract size could not be easily normalized in this population.

Although it was evident that the use of FCRs and Bark units reduced variance in the acoustic measurements, the question remained: Do these techniques also eliminate important information regarding articulatory movement? The results presented in Table 5 suggest that this is not the case. In comparison with triangular VSA, the use of an adapted FCR consistently improved the acoustic-perceptual relationship among speakers (accounting for an average of 6% more variance in men and 11% in women). This result was consistent with findings from previous studies that have compared FCR to triangular VSA (Lansford & Liss, 2014a; Sapir et al., 2010) and demonstrates the utility of using formant ratios as a measurement tool in other dialects. As hypothesized, the use of the Bark unit of frequency tended to increase the relationship between VSA measures and perceptual ratings (with a 6% increase in variance accounted for in men and 10% in women). When the FCR was used, there was a much smaller effect of using Bark units (an average of 5% increase in the female data, with no increase observed in the male data).

It is worth noting that both the original FCR and our adapted formula use measurements of only three corner vowels. In dialects with more corner vowels, the inclusion of additional formants in this ratio may benefit the measurement's validity. For example, Lansford and Liss (2014a) found that the FCR was able to account for approximately 15% more variance in intelligibility measurements than triangular VSA (when using the same three vowels). However, this difference reduced to just 3% when the quadrilateral vowel space was used—indicating that a set of four vowels may more adequately index the vowel dispersion of these U.S. speakers. In the case of NZE, the triangular shape of the dialect's vowel dispersion lends itself well to the three-vowel FCR and, for this reason, may have boosted the success of the measurement tool (Maclagan, 2009).

Overall, when making these interspeaker comparisons, it appears that formant ratios have the capacity to map more strongly to our perceptual impressions than vowel space measurements. However, in dialects with a more quadrilateral dispersion of vowels, the effect of Bark units on quadrilateral VSA (and the inclusion of more vowels in formant ratios) warrants further examination.

Perceptual Correlates

Findings of the current study suggest that small changes to listener prompts may affect the way perceptual ratings of dysarthria are distributed. Although it is apparent that the two measurements used in this study were highly correlated, the distribution of ratings meant that there was less variation among the "above average" scores of intelligibility. This was consistent with our hypothesis that ratings of intelligibility may not be as sensitive to mild speech disorder as ratings of speech precision. Indeed, ratings of speech precision tended to better separate the speakers with dysarthria from healthy controls. Across all metrics of vowel centralization, ratings of speech precision explained the most acoustic variance between speakers. Although the improvement was subtle, it is suggested that ratings of speech precision may capture changes in vowel centralization more successfully than ratings of intelligibility.

There are many ways to perceptually scale dysarthria severity that were not investigated in this study. For example, several previous studies have focused on comparisons of equal interval scales and direct magnitude estimates of speech disorder (e.g., Eadie & Doyle, 2002; Schiavetti, Metz, & Sitler, 1981; Zraick & Liss, 2000). These studies have suggested that listeners will not necessarily divide speech stimuli into intervals with an equal magnitude of change between them (i.e., the magnitude of change between a rating of 1 and 2 may be different from the change between 5 and 6).

The current investigation used VAS to rate dysarthria. Unlike equal interval scales, VAS do not force listeners to partition speech samples into categories—and may have allowed the listeners to better index differences in the magnitude of speakers' intelligibility and speech precision. In the current study, VAS enabled listeners to record their judgments quickly, with high reliability. The resultant ratings were able to account for up to 76% of the variance in vowel centralization measures—providing good evidence of their utility in measuring the speech signal. It is possible that direct magnitude estimates may also have produced results sensitive to acoustic change. Comparisons of direct magnitude estimates and VAS ratings should be explored in future work, particularly when large numbers of speech stimuli are being assessed. These comparisons should focus on the ability of the scales to index objective changes in the speech signal—rather than simply comparing the distributions of listener scores.

Limitations and Conclusions

In recommending changes in measurement procedures, careful consideration must be given to the generalizability of this study's results. Despite substituting the NZE [o:] vowel in place of the /u/ phoneme (which has been traditionally used in the dysarthria literature), the average midpoint VSA and FCR values in this study did not significantly differ from results presented in other large-scale studies of U.S. speakers with dysarthria (Lansford & Liss, 2014b; Sapir et al., 2010). For example, to compare the values obtained from the adapted FCR measure to values collected using the formula presented in Sapir et al. (2010), the average FCRs (generated from midpoint vowel formants) were examined. In the current study, speakers with dysarthria had an average FCR of 1.14 (SD = 0.12), whereas healthy speakers had an average FCR of 1.03 (SD = 0.07). These results lie directly between those reported by Lansford and Liss (2014b) and Sapir et al. (2010)—both of whom recruited speakers from a similar geographic region of the United States and used the original FCR formula. The adapted FCR produced averages for NZE speakers (both with and without dysarthria) that were within one standard deviation of the values reported in these studies. Evidently, there are considerable variations reported in VSA and FCR values as well as differences in midpoint vowel formant values of the /a/, /i/ and /u/ phonemes, among healthy speakers of U.S. English (Lansford & Liss, 2014b; Sapir et al., 2010; Turner et al., 1995). For this reason, it is difficult to determine how differences in the raw VSA and FCR values were influenced by the NZE dialect.

The current investigation found large differences in the acoustic perceptual relationships demonstrated by male and female speakers. Considerably stronger correlations were produced among female speakers, indicating that changes in their acoustic measurements were more closely related to listeners' perceptions of dysarthria. Given the smaller number of female participants, random sample variation may have played a role in producing this result. However, this study is not the first to find a stronger link between perceptual and acoustic vowel measures among female speakers (see Lansford & Liss, 2014a). It is possible that current methods of indexing vowel centralization might influence differences between the sexes. Both metrics

amplify F1/F2 changes differently depending on the magnitude of speakers' formants, and this may account for some of the differences in the acoustic-perceptual relationship across the sexes. Examining the same set of speakers over time may help elucidate how the relationship between vowel centralization metrics and perceptual measurements is affected by differences in the magnitude of baseline formant values.

In summary, this investigation found that changes in the methods used to assess vowel centralization and speech severity resulted in differences in the way these measurements indexed speech disorder, with some techniques more clearly distinguishing people with dysarthria from healthy controls. In addition, these changes resulted in variable strengths of relationship between acoustic and perceptual measures of dysarthria. Taken together, the techniques suggested in this study were able to double the amount of variance accounted for in the acoustic-perceptual relationship among male speakers. In women, the amount of variance accounted for increased from 49% to 76%. This demonstrates that the procedures chosen when taking these measurements have an important influence on a study's results. The ability to achieve stronger acoustic-perceptual relationships in individuals with dysarthria is vital in order to validate our acoustic measurements for clinical use. In future vowel space studies, it is recommended that researchers consider more flexible formant extraction points and different normalization procedures. Furthermore, it should be noted that perceptual measurements of dysarthria are not an inflexible standard, and the procedures we use to rate dysarthria should be carefully considered when any acoustic metrics are being assessed. Perceptual rating tasks that allow listeners to indicate that speech sounds impaired —even if the signal is intelligible—appear to be advantageous when indexing changes in mild speech disorder.

References

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1996). *Celex2* (LDC96L14). Philadelphia, PA: Linguistic Data Consortium.

Boersma, P., & Weenink, D. (2012). *Praat: Doing phonetics by computer*. Version 5.3.04. Retrieved from http://www.praat.org/

Clopper, C. G. (2009). Computational methods for normalizing acoustic vowel data for talker differences. *Language and Linguistics Compass, 3,* 1430–1442.

Cox, F. (2006). The acoustic characteristics of /hVd/ vowels in the speech of some Australian teenagers. *Australian Journal of Linguistics, 26,* 147–179.

Diehl, R. L., Lindblom, B., Hoemeke, K. A., & Fahey, R. P. (1996). On explaining certain male-female differences in the phonetic realization of vowel categories. *Journal of Phonetics, 24,* 187–208.

Eadie, T. L., & Doyle, P. C. (2002). Direct magnitude estimation and interval scaling of pleasantness and severity in dysphonic and normal speakers. *Journal of the Acoustical Society of America, 112,* 3014–3021.

Ferguson, S. H., & Kewley-Port, D. (2007). Talker differences in clear and conversational speech: Acoustic characteristics of vowels. *Journal of Speech, Language, and Hearing Research, 50,* 1241–1255.

Fletcher, A. R., McAuliffe, M. J., Lansford, K. L., & Liss, J. M. (2015). The relationship between speech segment duration and vowel centralization in a group of older speakers. *Journal of the Acoustical Society of America, 138,* 2132–2139.

Fromont, R., & Hay, J. (2008). ONZE Miner: The development of a browser-based research tool. *Corpora, 3,* 173–193.

Hillenbrand, J. M., Clark, M. J., & Nearey, T. M. (2001). Effects of consonant environment on vowel formant patterns. *Journal of the Acoustical Society of America, 109,* 748–763.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America, 97,* 3099–3111.

Kim, H., Hasegawa-Johnson, M., & Perlman, A. (2011). Vowel contrast and speech intelligibility in dysarthria. *Folia Phoniatrica et Logopaedica, 63,* 187–194.

Kim, Y., 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,* 417.

Lansford, K. L., & Liss, J. M. (2014a). Vowel acoustics in dysarthria: Mapping to perception. *Journal of Speech, Language, and Hearing Research, 57,* 68–80.

Lansford, K. L., & Liss, J. M. (2014b). Vowel acoustics in dysarthria: Speech disorder diagnosis and classification. *Journal of Speech, Language, and Hearing Research, 57,* 57–67.

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. *Journal of the Acoustical Society of America, 117,* 3879–3889.

Maclagan, M. (2009). Reflecting connections with the local language: New Zealand English*. *International Journal of Speech-Language Pathology, 11,* 113–121.

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,* 35–50.

Neel, A. T. (2008). Vowel space characteristics and vowel identification accuracy. *Journal of Speech, Language, and Hearing Research, 51,* 574–585.

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,* 114–125.

Schiavetti, N., Metz, D. E., & Sitler, R. W. (1981). Construct validity of direct magnitude estimation and interval scaling of speech intelligibility: Evidence from a study of the hearing impaired. *Journal of Speech, Language, and Hearing Research, 24,* 441–445.

Sheard, C., Adams, R. D., & Davis, P. J. (1991). Reliability and agreement of ratings of ataxic dysarthric speech samples with varying intelligibility. *Journal of Speech and Hearing Research, 34,* 285–293.

Sussman, J. E., & Tjaden, K. (2012). Perceptual measures of speech from individuals with Parkinson's disease and multiple sclerosis: Intelligibility and beyond. *Journal of Speech, Language, and Hearing Research, 55,* 1208–1219.

Tjaden, K., & Wilding, G. E. (2004). Rate and loudness manipulations in dysarthria: Acoustic and perceptual findings. *Journal of Speech, Language, and Hearing Research, 47,* 766–783.

Traunmüller, H. (1990). Analytical expressions for the tonotopic sensory scale. *Journal of the Acoustical Society of America, 88,* 97–100.

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, Language, and Hearing Research, 38,* 1001–1013.

Weismer, G., & Berry, J. (2003). Effects of speaking rate on second formant trajectories of selected vocalic nuclei. *Journal of the Acoustical Society of America, 113,* 3362–3378.

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,* 1–18.

Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., . . . Woodland, P. (2002). *The HTK-Book 3.2.* Cambridge, United Kingdom: Cambridge University Press.

Zraick, R. I., & Liss, J. M. (2000). A comparison of equal-appearing interval scaling and direct magnitude estimation of nasal voice quality. *Journal of Speech, Language, and Hearing Research, 43,* 979–988.

Appendix A

The Grandfather Passage

You wish to know all about my grand**fath**er. Well, he is nearly 9**3** years old, yet he still thinks as swiftly as ever. He dresses himself in an old, black frock coat, usually with several buttons missing. A long beard clings to his chin, giving those who observe him a pronounced feeling of the utmost respect. Twice **each** day he plays skillfully and with zest upon a small **org**an. Except in the winter when the snow or ice prevents, he slowly takes a **short** walk in the open air **each** day. We have often urged him to walk **more** and smoke less but he always **ans**wers, "Banana oil!" Grand**fath**er likes to be modern in his language.

Note. Syllables in bold indicate where the New Zealand point vowels were extracted from the reading passage.

Appendix B

Listener Rating Instructions

For the articulatory precision ratings, the following instructions were given: "In this experiment, you will rate people's speech precision. Precise speech sounds crisp, with clear and accurate enunciation. Some of the people you will hear have speech disorders which affect the precision of their speech. Your job is to judge each person's speech precision."

For the ratings of ease of understanding, the following instructions were given: "In this experiment, you will rate how easy it is to understand different speakers. Some of the people you will hear have speech disorders which affect how easy they are to understand. You will make your rating by placing a mark on a scale."