JSLHR

Research Article

Predicting Intelligibility Gains in Dysarthria Through Automated Speech Feature Analysis

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

Purpose: Behavioral speech modifications have variable effects on the intelligibility of speakers with dysarthria. In the companion article, a significant relationship was found between measures of speakers' baseline speech and their intelligibility gains following cues to speak louder and reduce rate (Fletcher, McAuliffe, Lansford, Sinex, & Liss, 2017). This study reexamines these features and assesses whether automated acoustic assessments can also be used to predict intelligibility gains.
Method: Fifty speakers (7 older individuals and 43 with dysarthria) read a passage in habitual, loud, and slow speaking modes. Automated measurements of long-term average spectra, envelope modulation spectra, and Mel-frequency cepstral coefficients were extracted from short segments of participants' baseline speech. Intelligibility gains were statistically modeled, and the predictive power of the baseline speech measures was assessed using cross-validation.
Results: Statistical models could predict the intelligibility gains of speakers they had not been trained on. The automated acoustic features were better able to predict speakers' improvement in the loud condition than the manual measures reported in the companion article.
Conclusions: These acoustic analyses present a promising tool for rapidly assessing treatment options. Automated measures of baseline speech patterns may enable more selective inclusion criteria and stronger group outcomes within treatment studies.

Behavioral speech modification is a primary focus of intervention aimed at improving intelligibility in speakers with dysarthria. Indeed, techniques involving increasing loudness and reducing speech rate have been promoted for speakers with a range of dysarthria etiologies (Fox & Boliek, 2012; Sapir et al., 2003; Van Nuffelen, De Bodt, Vanderwegen, Van de Heyning, & Wuyts, 2010). However, the efficacy of these strategies is not entirely clear. Not all speakers show improvement in intelligibility scores when cued to alter their speech (McAuliffe, Fletcher, Kerr, O'Beirne, & Anderson, 2017; Neel, 2009; Pilon, McIntosh, & Thaut, 1998; Tjaden & Wilding, 2004; Turner, Tjaden, & Weismer, 1995; Van Nuffelen, De Bodt, Wuyts, & Van de Heyning, 2009; Van Nuffelen et al., 2010), and it is not uncommon for treatment studies to fail to demonstrate intelligibility improvements across speaker groups (e.g., Lowit, Dobinson, Timmins, Howell, & Kröger, 2010; Mahler & Ramig, 2012). To promote speech therapy for speakers with dysarthria, we must become adept at identifying speakers who are likely to make positive treatment gains. This will allow us to better target our treatment strategies, enabling more selective inclusion criteria and stronger group outcomes within treatment studies.

This study is the second of two (Fletcher, McAuliffe, Lansford, Sinex, & Liss, 2017) that aim to investigate whether detailed measures of speakers' baseline speech can be used to predict their intelligibility gains following behavioral speech modification. At present, there are limited protocols for determining whether a treatment technique is appropriate for a given speaker. Researchers most often select speakers for treatment based on their Mayo system

[a]Department of Communication Disorders, University of Canterbury, Christchurch, New Zealand
[b]School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe
[c]School of Communication Science & Disorders, Florida State University, Tallahassee
[d]Department of Speech and Hearing Science, Arizona State University, Tempe
Correspondence to Annalise Fletcher: annalise.fletcher@canterbury.ac.nz

Editor-in-Chief: Krista Wilkinson
Editor: Jeannette Hoit
Received December 15, 2016
Revision received June 27, 2017
Accepted June 27, 2017
https://doi.org/10.1044/2017_JSLHR-S-16-0453

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

subtype (Darley, Aronson, & Brown, 1975) or neurogenic etiology, with participants with matching etiologies or subtypes grouped together in studies (e.g., placing all participants with Parkinson's disease and hypokinetic dysarthria within a single treatment group; Cannito et al., 2012; Lowit et al., 2010). However, there is little evidence to suggest that speakers who share a subtype or dysarthria etiology will respond similarly to treatment strategies (McAuliffe et al., 2017; Tjaden & Wilding, 2004; Van Nuffelen et al., 2010).

In the companion article (Fletcher et al., 2017), we demonstrated that perceptual measurements of speech severity, in addition to acoustic measurements of articulation rate and temporal variability, accounted for between 31% and 34% of the variation in intelligibility change across speakers. This supports the notion that measures of baseline speech features in dysarthria can be used to make predictions about how speakers might respond to treatment. However, the measures that best accounted for these intelligibility gains involved labor-intensive segmentation of the speech signal and large numbers of listeners to reliably quantify speech severity. The current study investigates whether automated measures of the speech signal can account for similar variability in speakers' intelligibility gains to those derived from the manually obtained acoustic/perceptual measures.

Automated Analyses of Dysarthric Speech

Recently, there has been considerable interest in the application of automated speech measures as a method of gathering faster, less invasive assessments of speakers' disease progression (for a review, see Bayestehtashk, Asgari, Shafran, & McNames, 2015). In the motor speech literature, several feature sets have been used to characterize the speech signal: long-term average spectra (LTAS), envelope modulation spectra (EMS), and Mel-frequency cepstral coefficients (MFCCs). Many perceptual qualities of speech (e.g., articulatory imprecision, hypernasality, and strained voicing) do not have simple acoustic correlates. Thus, although automated acoustic measures can provide detailed descriptions of the shape and composition of the acoustic waveform, there is no definitive one-to-one mapping of specific measurements of LTAS, MFCCs, and EMS to perceptual speech features. Despite these limitations, measures of LTAS, EMS, and MFCCs have obvious clinical potential. They can systematically quantify the speech signal in a manner that is easy to compute and does not require manual processing (e.g., Liss, LeGendre, & Lotto, 2010). Hence, combinations of these features can be used as a basis for modeling speech severity and rapidly assessing treatment options (Berisha, Utianski, & Liss, 2013).

LTAS have been commonly used in the speech disorder literature to index differences in voice quality and nasality across speech samples (Lowell, Colton, Kelley, & Hahn, 2011). LTAS provide a representation of the average spectral information contained in the speech signal across a relatively lengthy period (i.e., they provide information about the spectral content across whole phrases, rather than within specific phonemes). Tjaden, Sussman, Liu, and Wilding (2010) and Lowell et al. (2011) demonstrated that there are significant correlations between LTAS measures and perceptual ratings of dysarthria severity and voice disorder. Improvements in voice quality are usually demonstrated through a strengthening of lower-frequency components of the LTAS and a weakening of upper-frequency components (Cannito et al., 2012). For example, Tanner, Roy, Ash, and Buder (2005) observed that speakers with functional dysphonia had lower spectral means and standard deviations following behavioral therapy. There is also evidence that LTAS measures can be used to detect changes in nasality, with amplitudes around 250 Hz showing significant changes when speakers simulate hypernasality (de Boer & Bressmann, 2016).

Another promising tool in the automatic evaluation of dysarthria is the measures of EMS. EMS represent modulations that occur in the amplitude of the speech signal. Slow rate modulations in amplitude can provide information about individual's articulatory rate as well as any sudden changes in loudness or interruptions to the speech signal. Liss et al. (2010) explored measurements derived from EMS that were taken from a range of frequency bands within the speech signal. They found that these features were 95% accurate in classifying speakers with dysarthria from healthy controls on cross-validation. Furthermore, they demonstrated 67% accuracy in their ability to classify individuals into five speaker groups. These groups included four different dysarthria subtypes, in addition to a group of healthy control speakers. Liss et al. (2010) selected speakers into these four dysarthria groups because they exhibited the cardinal perceptual features thought to be associated with their subtype. Hence, their findings suggest that EMS measures may be particularly sensitive to perceptual differences associated with dysarthria etiology and subtype.

MFCCs provide the most widely used representations of the speech signal in automated speech recognition programs and are becoming increasingly common in analyses of speech disorders (Han, Chan, Choy, & Pun, 2006; Paja & Falk, 2012). Broadly speaking, MFCCs are used to capture information about the spectral structure of speech over time in a manner that approximates the way we perceive speech sounds. In the study of dysarthria, they are used with the aim of measuring subtle changes in the movement of articulators (Khan, Westin, & Dougherty, 2014). For example, Van Nuffelen, Middag, De Bodt, and Martens (2009) examined whether acoustic models derived from MFCCs could be used to estimate listener ratings of intelligibility. This study used acoustic models to compute the probability that each analysis frame would be aligned to a phoneme or phonological property. Their final models of intelligibility achieved correlations of up to .94 with listener intelligibility scores.

Summary and Aims of the Current Study

The current study used LTAS, EMS, and MFCC features to predict whether speakers would benefit from different speech modification strategies. Unlike the manual acoustic/perceptual measures presented in the companion article, these techniques do not require prior segmentation of phonemes and can be applied to relatively short speech samples. Thus, it is hoped that they may be more readily integrated as a speech assessment tool in clinical research. The specific purpose of the current study was to determine whether these automated measures could be used to model variation in speakers' intelligibility gains. The robustness of these statistical models was tested using cross-validation techniques. This approach enabled us to assess whether newly generated statistical models were generalizable to new groups of speakers. Manual acoustic/perceptual measurements from the companion article were also re-analyzed to determine their predictive power using the same cross-validation techniques. With this information, we compared the performance of the two speech assessment methods.

Method

Speakers and Speech Stimuli

Fifty speakers contributed speech recordings to this study (seven healthy older individuals and 43 with dysarthria). The intention was to be able to acoustically characterize a full spectrum of dysarthria features. Thus, the speakers had a range of dysarthria etiologies and speech severities. For full demographic details, see Table 1.

Procedures for speech recording are described in the companion article (Fletcher et al., 2017). "The Grandfather Passage" was used to elicit a sample of participants' baseline speech, as well as samples simulating two common treatment strategies. For the baseline condition, speakers were asked to read the passage in their everyday speaking voice after they had familiarized themselves with the passage. To create the treatment simulations, a magnitude scaling procedure was used to elicit louder and slower speech. The procedure for this is described in McAuliffe, Kerr, Gibson, Anderson, and LaShell (2014).

Procedure

To assess whether features from automated speech analyses could predict the intelligibility gains of these speakers, two sets of data were required for each speaker: (a) perceptual ratings of intelligibility gain in the loud and slow speaking conditions and (b) measurements of LTAS, EMS, and MFCC features from the baseline speech condition.

Perceptual Ratings of Intelligibility Gain

Perceptual ratings of intelligibility gain were collected in the companion study and are described in detail in the article (Fletcher et al., 2017). Briefly, to determine intelligibility gains in the loud and slow conditions, 18 listeners rated speakers' intelligibility on a visual analogue scale. In each trial, the listeners were presented with three matching phrases, extracted from "the Grandfather Passage," that were produced in the baseline, loud, and slow conditions by one of the study's speakers. Listeners were blinded to the identity of the three conditions. The presentation of these phrases was randomized in each trial, and the listeners were not informed about the nature of the speech modifications. In each trial, listeners were instructed to rate how easy the speech was to understand. Although the listeners heard different sets of phrases for different speakers, all speech stimuli were between 11 and 14 syllables in length. The distance that listeners placed between the baseline and treatment tokens was used to calculate two intelligibility gain indices for each speaker: one for change in the slow condition and one for change in the loud condition.

Measurements of Automated Acoustic Features From the Baseline Speech Condition

To automatically extract acoustic features of speakers' baseline speech, three sets of features were obtained via MATLAB scripts (as previously reported in Berisha, Sandoval, Utianski, Liss, & Spanias, 2013; Liss et al., 2010; Wisler, Berisha, Liss, & Spanias, 2014). All features were extracted from an identical phrase within speakers' baseline speech sample. The phrase "he slowly takes a short walk in the open air each day" was chosen for this purpose. The phrase was specifically selected because the 50 baseline recordings were free from reading errors and there were no significant nonspeech sounds or periods of silence produced within the phrase. This helped ensure consistency in the acoustic analyses. The feature sets used in the automatic analysis of the phrase are described in the following sections.

LTAS

To extract the LTAS values, the speech signal was passed through an octave filter, breaking it into nine separate bands. The center frequencies of these bands were 30, 60, 120, 240, 480, 960, 1920, 3840, and 7680 Hz. For each of the nine octave bands and the full signal, the data were framed using a 20-ms rectangular window with no overlap to calculate (a) the normalized average root-mean-square (RMS) energy, (b) the RMS energy range, (c) the normalized RMS energy range, (d) skew, (e) kurtosis, as well as the standard deviation of RMS energy normalized relative to both (f) the total RMS energy and (g) the RMS energy in each band (not applicable in the analysis of the full signal). We also extracted (h) the pairwise variability of RMS energy between successive frames, (i) the mean of the framed RMS energies, and (j) the normalized mean of the framed RMS energies. This produced 99 LTAS features. Measures of distribution (including standard deviation, skew, and kurtosis) are frequently used to quantify differences in LTAS (Tanner et al., 2005). The procedures used in this study allowed us to examine the energy range and distribution in the full speech signal, while also focusing

Table 1. Demographic information for speakers with dysarthria.

Gender	Age	Medical etiology	Severity of disorder
F	46	Brain tumor	Mild–moderate
M	58	Brainstem stroke	Moderate
M	56	Cerebellar ataxia	Mild
F	69	Cerebral palsy	Severe
M	60	Cerebral palsy	Severe
F	68	Freidreich’s ataxia	Mild
F	47	Huntington’s disease	Moderate–severe
M	55	Huntington’s disease	Severe
M	43	Hydrocephalus	Severe
F	53	Multiple sclerosis	Mild–moderate
F	60	Multiple sclerosis	Moderate–severe
F	79	Parkinson’s disease	Mild
M	76	Parkinson’s disease	Mild
M	77	Parkinson’s disease	Mild
M	67	Parkinson’s disease	Mild
F	83	Parkinson’s disease	Mild
M	68	Parkinson’s disease	Mild
M	89	Parkinson’s disease	Mild
M	58	Parkinson’s disease	Mild
M	73	Parkinson’s disease	Mild
M	79	Parkinson’s disease	Mild
M	69	Parkinson’s disease	Mild
M	68	Parkinson’s disease	Mild
F	70	Parkinson’s disease	Mild–moderate
M	67	Parkinson’s disease	Mild–moderate
M	71	Parkinson’s disease	Mild–moderate
F	73	Parkinson’s disease	Mild–moderate
M	65	Parkinson’s disease	Mild–moderate
M	75	Parkinson’s disease	Moderate
M	79	Parkinson’s disease	Moderate
M	71	Parkinson’s disease	Moderate
M	69	Parkinson’s disease	Moderate
M	81	Parkinson’s disease	Moderate–severe
M	77	Parkinson’s disease	Moderate–severe
F	67	Progressive supranuclear palsy	Mild
M	64	Spinocerebellar ataxia	Severe
M	72	Stroke	Severe
F	48	Traumatic brain injury	Mild–moderate
M	55	Traumatic brain injury	Mild–moderate
M	60	Traumatic brain injury	Moderate
M	47	Traumatic brain injury	Severe
M	53	Undetermined neurological disease	Moderate
F	45	Wilson’s disease	Mild

Note. F = female; M = male.

on specific sections of the LTAS by separately considering the energy contained within the nine octave bands.

EMS

Before obtaining the EMS, speech recordings were filtered into nine frequency bands with center frequencies of 30, 60, 125, 250, 500, 1000, 2000, 4000, and 8000 Hz. Amplitude envelopes were taken from the nine bands as well as the full speech signal. Low-pass filters with a cutoff of 30 Hz were then applied to the amplitude envelopes to capture the slower changes in amplitude that occur across words and phrases. Fourier analyses were used to quantify the temporal modulations in the signal. Six EMS metrics were computed for each of the nine bands and the full signal: (a) peak frequency, (b) peak amplitude, (c) energy in the spectrum from 3 to 6 Hz, (d) energy in spectrum from 0 to 4 Hz, (e) energy in spectrum from 4–10 Hz, and (f) energy ratio between 0–4 Hz band and 4–10 Hz band. Note that (c), (d), and (e) are normalized by the total energy in the EMS between 0 and 10 Hz. This resulted in a total of 60 EMS features. Liss et al. (2010) provides further explanation as to why these six metrics were developed for the analysis of speech prosody.

MFCCs

The MFCCs were calculated using a filter bank approach described in Vergin, O’shaughnessy, and Farhat (1999), where the speech signal was filtered into 39 frequency bands distributed approximately evenly along the Mel scale. The first 20 filters were linearly spaced between 0 and 1000 Hz, whereas the next 19 were logarithmically spaced between 1 and 22.05 kHz. For each of the 39 Mel-filtered signals, the data were framed using 20-ms Hamming windows with 10-ms overlap, and the log energy was calculated. Taking the inverse discrete cosine transform of the 39 filtered log energy values for each frame yields the set of 39 MFCCs for that frame. Within each of these 39 MFCCs, six different statistics were computed: (a) mean, (b) standard deviation, (c) range, (d) pairwise variability, (e) skew, and (f) kurtosis. This resulted in 234 MFCC features. Many studies that investigate automatic speech recognition extract only the first 13 coefficients to represent the envelope of the short-term power spectra. By extracting a larger number of MFCCs and examining their distribution, we attempt to gain additional spectral details that could illuminate individual differences in speech production.

Measurements of Manual Acoustic/Perceptual Variables

The aim of this study was to predict speakers' intelligibility gains using automatically generated baseline speech features and compare these models to predictions made with manual measurements of the speech signal. The extraction of these manual features is reported in our companion article (Fletcher et al., 2017). The manual features included a perceptual rating of speech severity and acoustic measurements of speakers' articulation rates, formant centralization ratios, the amplitudes of their first harmonics relative to second, cepstral peak prominences, the standard deviation of their pitch and amplitude from across the passage reading, and the pairwise variability index of their vowels. These data were reanalyzed in the current study and used to create new statistical models. These statistical models assess how well the manual features predict the performance of untrained speakers (i.e., their predictive power on cross-validation). The statistical procedures are described in the following section.

Statistical Analyses

Four regression models were developed. Models 1 and 2 predicted the degree that speakers changed their intelligibility in the slow condition relative to their baseline speech sample. The first model selected independent variables from the eight manual speech measures reported in the companion study. The second model selected its independent variables from the full range of automated acoustic feature sets (including measures of MFCCs, EMS, and LTAS). Models 3 and 4 predicted the degree that speakers changed their intelligibility in the loud condition relative to their baseline speech sample. Model 3 selected independent variables from the eight manual speech measures. Model 4 selected variables from the full range of automated acoustic features. Each model's predictive power was assessed by determining the correlations between the intelligibility gain predicted by the model's output and the real intelligibility gains made by the speakers.

The large number of automated features extracted from speakers' baseline speech (a combined total of 393 features per speaker) meant that standard stepwise regression methods needed to be applied with caution in Models 2 and 4. For example, a forward stepwise regression (with α set to p = .05) would likely continue adding variables until it overfits the perceptual data. Hence, a cross-validation procedure was applied to determine the total number of features to be included in all models. This cross-validation procedure provided a measurement of the amount of variation that the statistical models could predict in the intelligibility gains of speakers they had not been trained on. A 10-fold cross-validation procedure was employed, where the speakers were divided at random into 10 equal groups. Cross-validation was achieved by training models on 90% of the speakers and testing their predictive power on the remaining 10% (test speakers). This process was then repeated a further nine times. On each repetition, a different set of nine speaker groups was used to train the model, and the model was tested on the remaining group. The predictive power was then averaged across the 10 repetitions to determine the models' performance.

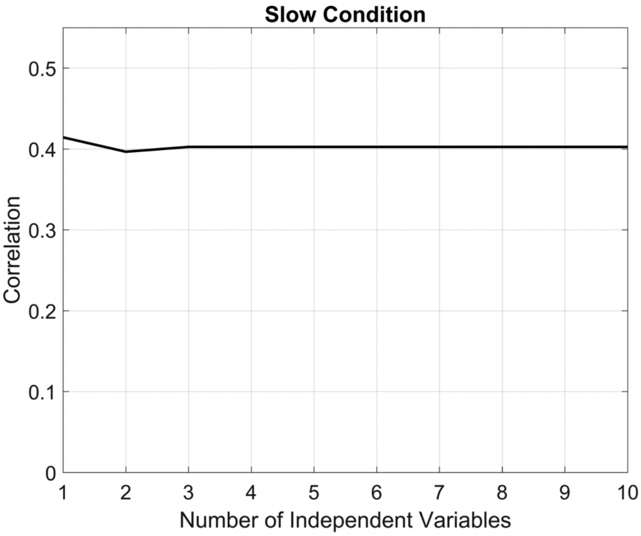
This cross-validation procedure was used to test the performance of the models each time a new variable was added to the forward stepwise regressions. For example, initially only one independent variable was selected for each model. The selection of this variable was based on its correlation with the intelligibility gains of the speakers it was trained on. Because there were 10 different renditions of the model every time a variable was added (as it was trained 10 times on different combinations of speakers), a different variable could be selected for the model each time, depending on the training data. The performance of the 10 renditions was averaged to provide an estimate of how well the model could account for variations in intelligibility gain when only one independent variable was used. This process began anew in the second step of the forward regression, using the same 10 groups of speakers to train and test the model when it contained two independent variables (provided the second variable accounted for statistically significant additional variation, $p < .05$). Testing continued with up to 10 independent variables allowed in each of the models. The results of this procedure were used to determine at which point in the forward regression the predictive power of the cross-validated model was highest.

Results

Stepwise Regression to Predict Intelligibility Change in the Slow Condition

Figures 1 and 2 detail the relationship between participants' baseline speech features and their intelligibility gains in the slow speaking condition across different statistical models. The y-axis shows the correlation between

Figure 1. The predictive power of manual acoustic/perceptual measurements. The *y*-axis depicts the correlation between the intelligibility gain predicted by the models' output and the true intelligibility gain indices measured following cues to reduce rate. Average correlations are generated for each step of the forward regression (as shown along the *x*-axis).



the performance predicted by the statistical models and the true intelligibility gains made by the test speakers. This relationship is evaluated at each step in the forward regression model building process (up until a total of 10 variables had been added). The figures demonstrate the models' performance on speakers that they have not been exposed to or trained on. As described in the methods, this value is an average of 10 renditions of the stepwise regression, each tested on different speakers.

Figure 2. The predictive power of automated acoustic measurements. The *y*-axis depicts the correlation between the intelligibility gain predicted by the models' output and the true intelligibility gain indices measured following cues to reduce rate. Average correlations are generated for each step of the forward regression (as shown along the *x*-axis).

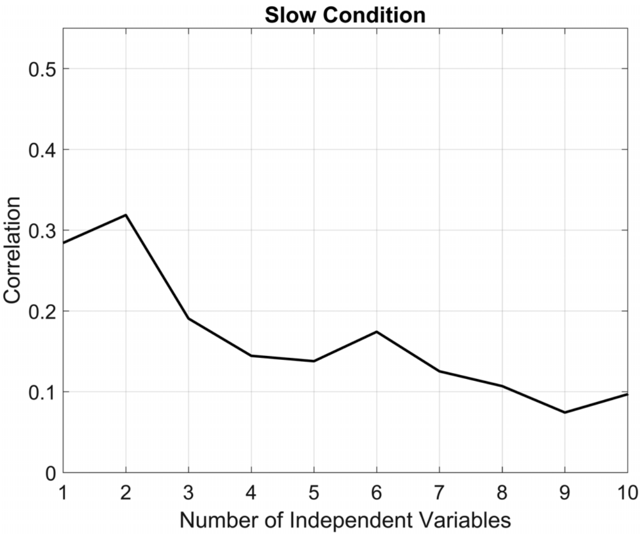


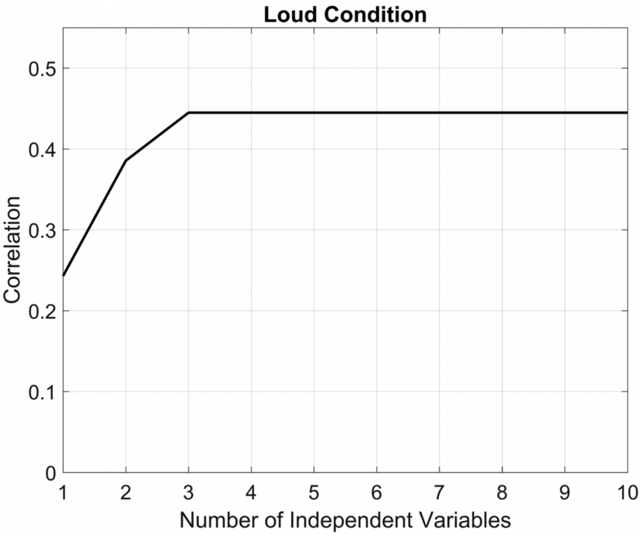
Figure 1 shows the performance of the eight manual variables (a perceptual rating of speech severity and acoustic measurements of speakers' articulation rates, formant centralization ratios, the amplitudes of their first harmonics relative to second, cepstral peak prominences, the standard deviation of their pitch and amplitude, and the pairwise variability index of their vowels). This figure demonstrates that, as the number of independent variables in the models increases, there is almost no change in the models' accuracy in predicting the intelligibility gains of the test speakers. Three steps into the forward stepwise regression, the performance of the models remains static. This demonstrates that the stepwise regression can no longer find any manual variables that would account for further statistically significant variation in intelligibility gains of any speakers it has been trained on (at p = .05). Overall, the models achieve their highest accuracy in predicting the performance of the test speakers when only one independent variable is included. When one variable is included, there is a correlation of .41 between the test speakers' true intelligibility gain indices and the average predicted by the models, indicating that the manual variables can account for approximately 17% of the variation in speakers' intelligibility changes.

Figure 2 illustrates the performance of the 393 automated acoustic features. When the number of independent variables increases above two, there is a sharp decline in the models' accuracy in predicting the intelligibility gains of the test speakers. This demonstrates that overfitting is occurring. The stepwise regression continues to select independent variables that account for significant variation in the intelligibility gains of the speakers it is trained on. However, it becomes less accurate in predicting the intelligibility gains of new speakers. This suggests that the additional independent variables have begun to describe small variations in intelligibility gains that are specific to this group of participants and do not reflect patterns in the larger population. Overall, the models achieve their highest accuracy in predicting the performance of the test speakers when two independent variables are included. At this point, there is an average correlation of .32 between the test speakers' true intelligibility gain indices and those predicted by the models, indicating that the manual variables can account for approximately 10% of the variation in speakers' intelligibility changes.

Stepwise Regression to Predict Intelligibility Change in the Loud Condition

Figures 3 and 4 detail the relationship between participants' baseline speech features and their intelligibility gains in the loud speaking condition. Figure 3 shows the performance of statistical models based on the eight

Figure 3. The predictive power of manual acoustic/perceptual measurements. The *y*-axis depicts the correlation between the intelligibility gain predicted by the models' output and the true intelligibility gain indices measured following cues to speak loud. Average correlations are generated for each step of the forward regression (as shown along the *x*-axis).



manual variables measured in our companion article. The accuracy of the models in predicting speakers' intelligibility gain indices increases until three independent variables are included. After this point, the performance of the models remains static. This demonstrates that the stepwise regression can no longer find any manual variables that would account for further statistically significant variation in intelligibility gains of any speakers it has been trained on (at p = .05). Overall, the models achieve their highest accuracy in predicting the performance of the test speakers when three independent variables are included. At this point, there is a correlation of .44 between the test speakers' true intelligibility gain indices and those predicted by the models, indicating that the manual variables can account for approximately 19% of the variation in speakers' intelligibility changes.

Figure 4. The predictive power of automated acoustic measurements. The *y*-axis depicts the correlation between the intelligibility gain predicted by the models' output and the true intelligibility gain indices measured following cues to speak loud. Average correlations are generated for each step of the forward regression (as shown along the *x*-axis).

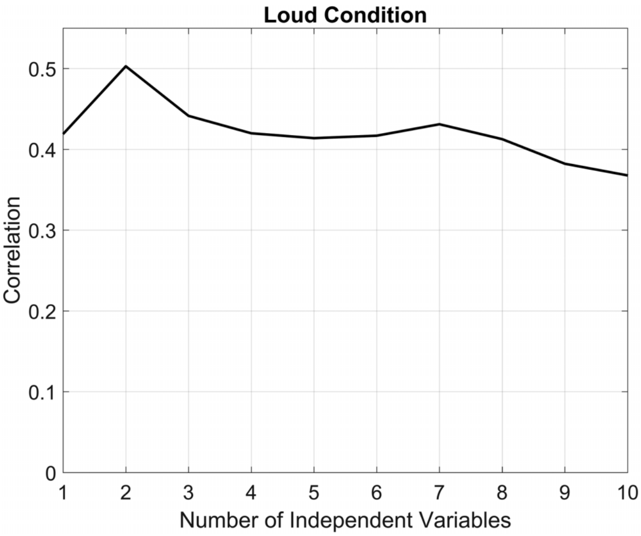


Figure 4 illustrates the performance of the 393 automated acoustic features. When the number of independent variables increases above two, overfitting occurs, and the models become less accurate in predicting the intelligibility gains of new speakers. Overall, the models achieve their highest accuracy in predicting the performance of the test speakers when two independent variables are included. At this point, there is a correlation of .50 between the test speakers' true intelligibility gain indices and those predicted by the models, indicating that the manual variables can account for approximately 25% of the variation in speakers' intelligibility changes.

Final Models of Intelligibility Gain

The final regression analyses were conducted using the full data set to model speakers' intelligibility gain indices. Results from the cross-validation procedure were used to determine the appropriate number of independent variables in each of the four models.

Predicting Intelligibility Gains in the Slow Condition

A one-step forward regression was conducted using the eight manual variables reported in Fletcher et al. (2017). The independent variable that showed the greatest correlation with the intelligibility gain indices was the first to be selected into the model. For predicting intelligibility gains in the slow condition, this variable was the perceptual rating of speakers' speech severity.

A two-step forward regression was conducted using the automatic acoustic feature sets. The first two variables to emerge from this regression were as follows: (a) the standard deviation extracted from calculations of the 34th MFCC from across the speech sample and (b) the amount of skewness in the distribution (from across the speech sample) of the 14th MFCC. Both variables were negatively associated with intelligibility gains.

Predicting Intelligibility Gains in the Loud Condition

A forward regression was conducted using the eight manual variables reported in Fletcher et al. (2017). Consistent with previous findings in the companion article, when the full data set was included in this regression, only two independent variables reached the threshold for statistical significance. The independent variables that showed the greatest correlation with the intelligibility gain indices in the loud condition were selected in the following order: (a) the articulation rate and (b) the perceptual rating of speakers' speech severity. A two-step forward regression was also conducted using the automatic acoustic feature sets. The first two variables to emerge from the regression were as follows: (a) the overall range observed in the measurements of the 14th MFCC and (b) a measurement of kurtosis in the 25th MFCC. Both measures were negatively associated with intelligibility gains.

Discussion

This current study explored whether automated acoustic analyses of baseline speech could be used to predict speakers' intelligibility gains following common treatment strategies. Specifically, this study focused on the performance of models on cross-validation. Cross-validation procedures were used to assess how accurately statistical models could predict the intelligibility gains of new groups of speakers. Overall, this study found that both the automated feature sets and the manual measures taken from Fletcher et al. (2017) predicted the intelligibility gains of speakers they had not been trained on. However, the variation they accounted for in the loud and slow speaking conditions was different. When speakers were prompted to speak louder, the statistical models were more accurate in predicting their intelligibility gains. The outcome of the cross-validation process was used to determine an appropriate method of model selection using the automated acoustic feature sets. The cross-validation procedure and the final models that resulted from these analyses are discussed in turn.

Power to Predict Intelligibility Gains

Cross-validation revealed that the models built using manual baseline measures and those built using the automated feature sets varied in their ability to predict the intelligibility gains of speakers they had not been trained on. The automated measures showed a clear reduction in their cross-validated performance after two features had been added to the regression. This suggested that the model had been overfitted. Overfitting occurs when dependent variables try to model random noise in the data set (e.g., fluctuations in speakers' intelligibility gains that are completely unrelated to their baseline speech). Allowing a model to choose from 393 features increases the likelihood that overfitting will occur because there is a greater chance that a feature will account for random—but statistically significant—variance in intelligibility gains. Overfit models generally have poor accuracy in predicting the outcomes of new groups of speakers, as demonstrated in Figures 2 and 4. When the manual measures were trained on 90% of the data points, they tended to stop adding features after two or three steps of the forward regression were completed and did not display the same tendency toward overfitting.

Overall, both feature sets demonstrated a stronger ability to predict intelligibility gains when speakers were cued to speak louder. The automated feature set demonstrated the strongest predictive power in this condition, accounting for 25% of the variance in speakers' intelligibility gains. This suggests that additional acoustic information—beyond the manual acoustic/perceptual measures used in Fletcher et al. (2017)—is important for determining the effectiveness of cues to speak louder. In contrast, the manual measures from Fletcher et al. (2017) were better able to predict speakers' intelligibility gains when cued to speak slower. In this case, the final model centered on measurements of baseline speech severity. Thus, it appeared that no single automated measurement of the acoustic speech signal accounted for more information than listeners' perceptual impressions.

Final Models: Which Features Best Predict Intelligibility Gains?

Forward stepwise regression was completed to identify which baseline speech factors were most closely related to intelligibility gains. Models of intelligibility gain that were based on the manual acoustic/perceptual measurements converged on the same independent variables as reported in the companion article (which utilized a backward stepwise regression approach). The relationship between these manual perceptual/acoustic measurements and speakers' intelligibility gains is discussed in detail in the companion article.

When using the manual measures, cross-validation results suggest that only one independent variable was able to predict intelligibility gain in the slow condition: the baseline perceptual rating of speech severity. In the loud condition, three independent variables provided statistically significant information about intelligibility gains. However, this was only true for some of the training groups. When examining the entire data set, the addition of a third variable in the loud condition did not improve the model; only speakers' baseline articulation rates and perceptual rating of speech severity were statistically significant predictors of intelligibility gain. This suggests that intelligibility gains produced by some speaker subgroups may be more accurately modeled using different baseline speech variables. Indeed, it is likely that further analysis of certain dysarthria subgroups could produce more effective, fine-tuned models of intelligibility gain. However, larger groups of speakers are needed to test this hypothesis.

Several automated acoustic measures demonstrated a strong relationship with speakers' intelligibility gains. When these features were used as independent variables, both final models selected measures derived from speakers' MFCC values. This presents a major challenge for interpreting these models because MFCCs are notoriously difficult to relate to perceptual characteristics of speech. This challenge stems from the fact that, although MFCCs are used to describe the shape of the spectrum, individual MFCCs cannot be tied any specific frequency or region of frequencies. So, although we know generally that higher MFCCs are representative of more rapid deviations in the log energy spectrum across frequency, understanding exactly what characteristics of the speech are captured by the 12th coefficient, but not by the 11th coefficient, is next to impossible. As a result, our efforts to interpret the perceptual attributes reflected by specific MFCC features are limited to broad conjectures, and we do not attempt to interpret why specific coefficients were selected over others.

Our models did not suggest the MFCCs themselves were relevant in determining whether speakers would produce intelligibility gains. Instead, it was the variation in the distribution of MFCCs that was most important. For example, in the slow condition, the standard deviation of the 34th MFCC had the strongest association with speakers' intelligibility gains. This likely indicates that the amount of articulatory movement from one phoneme to another is important (rather than the average value of the spectrum across the speech sample).

We also see evidence that higher-order MFCC features may be more important than lower MFCCs. Returning to the same example, the 34th MFCC represents rapid variations in the spectral structure with respect to frequency. This is likely to be indicative of smaller changes in the movement of the vocal tract. For example, it is more likely to represent small changes in vowel formant values than the broad differences produced when moving between a vowel and a consonant sound. We hypothesize that differences in these fine-grained movements of the vocal tract may be better able to index a speakers' dysarthria severity.

The second feature used to model intelligibility changes in the slow condition was skew in the distribution of the 14th MFCC. The 14th MFCC describes broader variations in the shape of the spectrum. One hypothesis for the occurrence of increased skew is that the precise articulation of particular phonetic targets may lead to outlying positive values at certain MFCCs. This would positively skew their distribution. Because it may be more challenging for speakers with dysarthria to produce these precise articulatory movements, we would expect this variable to also provide information about the severity of a person's dysarthria. We know that both the standard deviation of the 34th MFCC and the skewness of the 14th MFCC were negatively associated with intelligibility gains. These results are congruous with models of the manual acoustic/perceptual measures, which suggested that speakers who were perceived to have greater dysarthria severity produced larger intelligibility gains in the slow condition.

Intelligibility gains in the loud condition were also predicted by two MFCC measures, indicating the importance of articulatory information. First, the overall range of the 14th MFCC was selected. The MFCC range will be largest when a speaker has two analysis frames that contain spectrums that are starkly different in shape. Differences in spectrum shape could come about in a variety of ways. However, we would expect the largest differences to be occurring between vowels and consonants. For example, certain fricative sounds in the speech sample can be produced with intense, high-frequency energy. This would result in a distinctly different spectral pattern, a pattern that is likely to represent a sharper consonant sound. It is possible that speakers who produce larger ranges of movement (i.e., from a narrow constriction of the vocal tract to an unoccluded, open vowel) might exhibit a larger MFCC range. Larger ranges were negatively associated with intelligibility gains in the loud condition.

The second measure added to the model described the level of variation in the 25th MFCC, as measured through kurtosis. As with the 34th MFCC, we hypothesize that this MFCC is more representative of the fine-grained articulatory movements discussed previously. Hence, we would again hypothesize that this measure might be a strong indicator of overall dysarthria severity. This finding is congruous with models of the manual acoustic/perceptual measures, which suggested that speakers who were perceived to have greater dysarthria severity produced larger intelligibility gains in the loud condition. Differences in manual measurements of articulation rates were also a significant predictor of intelligibility gains in the loud condition, but it is unclear exactly how these are associated with speakers' MFCC variables.

Conclusions

Models of intelligibility gain accounted for a maximum of 25% of the variance in intelligibility gains in the speakers they were tested on. Hence, it appears that there could be considerably more factors, unmeasured in this study, that affect how speakers respond to speech modification strategies. This finding is not surprising. We know that people with dysarthria implement cues to speak louder and reduce rate in different ways (Tjaden & Wilding, 2004), and measures of cognitive ability, fatigue, depression, and self-efficacy are all believed to affect the success of treatment strategies (Fletcher & McAuliffe, 2017). Hence, although information about a person's baseline speech may convey valuable information about whether a treatment strategy is appropriate, it cannot entirely predict to what degree they will change their speech patterns and how listeners will perceive these changes.

There are also limitations inherent in relying on a single dependent variable to index speakers' intelligibility gains. This study analyzed ratings of intelligibility gain derived from a visual analogue scale. Although these ratings appeared to be reasonably sensitive to changes in speech patterns, listeners' preferences for different speech samples were not entirely consistent and reliable. This issue of listener subjectivity is discussed further in the companion article. More objective intelligibility measures based on orthographic transcription are prone to ceiling effects and may be less sensitive to some of the speech changes implemented by speakers with mild dysarthria (Sussman & Tjaden, 2012). However, they would allow us to elucidate whether some of the negative intelligibility changes observed in our data were reflective of a listener bias against slower or less natural sounding speech patterns, rather than a true reduction in the listeners' ability to understand the speaker. Hence, these measures could be a valuable addition to future studies.

Summary

In summary, automated acoustic features can be used as an indicator of how different speakers' intelligibility changes in response to common speech modifications. Furthermore, this information has the considerable benefit of being automatically extracted without any manual checking or subjective judgments from researchers. Objective diagnostic information is important for researchers wanting to develop more specific inclusion criteria for treatment studies. Furthermore, as automated acoustic measures continue to be incorporated into more user-friendly applications, these data may also be used to help provide recommendations for clinicians when choosing between different treatment programs. However, further development of the models presented in this study is required. Ideally, these models would benefit from training on larger groups of speakers. The inclusion of more data points in model training is likely to improve the cross-validated accuracy of models generated with three, four, or five variables. This would allow more confidence in their application to new groups.

Acknowledgments

This work was supported by a Fulbright New Zealand Graduate Award, granted to Annalise R. Fletcher.

References

Bayestehtashk, A., Asgari, M., Shafran, I., & McNames, J. (2015). Fully automated assessment of the severity of Parkinson's disease from speech. *Computer Speech & Language, 29*(1), 172–185.

Berisha, V., Sandoval, S., Utianski, R., Liss, J., & Spanias, A. (2013). *Selecting disorder-specific features for speech pathology fingerprinting*. Paper presented at the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, Canada.

Berisha, V., Utianski, R., & Liss, J. (2013). *Towards a clinical tool for automatic intelligibility assessment*. Paper presented at the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, Canada.

Cannito, M. P., Suiter, D. M., Beverly, D., Chorna, L., Wolf, T., & Pfeiffer, R. M. (2012). Sentence intelligibility before and after voice treatment in speakers with idiopathic Parkinson's disease. *Journal of Voice, 26*(2), 214–219.

Darley, F. L., Aronson, A. E., & Brown, J. R. (1975). *Motor speech disorders* (Vol. 304). Philadelphia, PA: Saunders.

de Boer, G., & Bressmann, T. (2016). Application of linear discriminant analysis to the long-term averaged spectra of simulated disorders of oral–nasal balance. *The Cleft Palate-Craniofacial Journal, 53*(5), e163–e171.

Fletcher, A., & McAuliffe, M. (2017). Examining variation in treatment outcomes among speakers with dysarthria. *Seminars in Speech and Language, 38*(03), 191–199.

Fletcher, A., McAuliffe, M., Lansford, K., Sinex, D., & Liss, J. (2017). Predicting intelligibility gains in individuals with dysarthria from baseline speech features. *Journal of Speech, Language, and Hearing Research, 60,* 3043–3057. https://doi.org/10.1044/2016_JSLHR-S-16-0218

Fox, C. M., & Boliek, C. A. (2012). Intensive voice treatment (LSVT LOUD) for children with spastic cerebral palsy and dysarthria. *Journal of Speech, Language, and Hearing Research, 55*(3), 930–945.

Han, W., Chan, C.-F., Choy, C.-S., & Pun, K.-P. (2006). *An efficient MFCC extraction method in speech recognition*. Paper presented at the 2006 IEEE International Symposium on Circuits and Systems, 2006 (ISCAS 2006, Proceedings), Kos, Greece.

Khan, T., Westin, J., & Dougherty, M. (2014). Classification of speech intelligibility in Parkinson's disease. *Biocybernetics and Biomedical Engineering, 34*(1), 35–45.

Liss, J. M., LeGendre, S., & Lotto, A. J. (2010). Discriminating dysarthria type from envelope modulation spectra. *Journal of Speech, Language, and Hearing Research, 53*(5), 1246–1255.

Lowell, S. Y., Colton, R. H., Kelley, R. T., & Hahn, Y. C. (2011). Spectral- and cepstral-based measures during continuous speech: Capacity to distinguish dysphonia and consistency within a speaker. *Journal of Voice, 25*(5), e223–e232.

Lowit, A., Dobinson, C., Timmins, C., Howell, P., & Kröger, B. (2010). The effectiveness of traditional methods and altered auditory feedback in improving speech rate and intelligibility in speakers with Parkinson's disease. *International Journal of Speech-Language Pathology, 12*(5), 426–436.

Mahler, L. A., & Ramig, L. O. (2012). Intensive treatment of dysarthria secondary to stroke. *Clinical Linguistics & Phonetics, 26*(8), 681–694.

McAuliffe, M. J., Fletcher, A. R., Kerr, S. E., O'Beirne, G. A., & Anderson, T. (2017). Effect of dysarthria type, speaking condition, and listener age on speech intelligibility. *American Journal of Speech-Language Pathology, 26*(1), 113–123.

McAuliffe, M. J., Kerr, S. E., Gibson, E. M., Anderson, T., & LaShell, P. J. (2014). Cognitive–perceptual examination of remediation approaches to hypokinetic dysarthria. *Journal of Speech, Language, and Hearing Research, 57*(4), 1268–1283.

Neel, A. T. (2009). Effects of loud and amplified speech on sentence and word intelligibility in Parkinson disease. *Journal of Speech, Language, and Hearing Research, 52*(4), 1021.

Paja, M. O. S., & Falk, T. H. (2012). *Automated dysarthria severity classification for improved objective intelligibility assessment of spastic dysarthric speech*. Paper presented at the 13th Annual Conference of the International Speech Communication Association, Portland, OR.

Pilon, M. A., McIntosh, K. W., & Thaut, M. H. (1998). Auditory vs visual speech timing cues as external rate control to enhance verbal intelligibility in mixed spastic ataxic dysarthric speakers: A pilot study. *Brain Injury, 12*(9), 793–803.

Sapir, S., Spielman, J., Ramig, L. O., Hinds, S. L., Countryman, S., Fox, C., & Story, B. (2003). Effects of intensive voice treatment (the Lee Silverman Voice Treatment [LSVT]) on ataxic dysarthria: A case study. *American Journal of Speech-Language Pathology, 12*(4), 387–399.

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*(4), 1208–1219.

Tanner, K., Roy, N., Ash, A., & Buder, E. H. (2005). Spectral moments of the long-term average spectrum: Sensitive indices of voice change after therapy? *Journal of Voice, 19*(2), 211–222.

Tjaden, K., Sussman, J. E., Liu, G., & Wilding, G. (2010). Long-term average spectral (LTAS) measures of dysarthria and their relationship to perceived severity. *Journal of Medical Speech-Language Pathology, 18*(4), 125–133.

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.

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.

Van Nuffelen, G., De Bodt, M., Vanderwegen, J., Van de Heyning, P., & Wuyts, F. (2010). Effect of rate control on speech production and intelligibility in dysarthria. *Folia Phoniatrica et Logopaedica, 62*(3), 110–119.

Van Nuffelen, G., De Bodt, M., Wuyts, F., & Van de Heyning, P. (2009). The effect of rate control on speech rate and intelligibility of dysarthric speech. *Folia Phoniatrica et Logopaedica, 61*(2), 69–75.

Van Nuffelen, G., Middag, C., De Bodt, M., & Martens, J. P. (2009). Speech technology-based assessment of phoneme intelligibility in dysarthria. *International Journal of Language & Communication Disorders, 44*(5), 716–730.

Vergin, R., O'shaughnessy, D., & Farhat, A. (1999). Generalized Mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition. *IEEE Transactions on Speech and Audio Processing, 7*(5), 525–532.

Wisler, A., Berisha, V., Liss, J., & Spanias, A. (2014). *Domain invariant speech features using a new divergence measure*. Paper presented at the 2014 IEEE Spoken Language Technology Workshop (SLT), Lake Tahoe, NV.