Acoustic Measures of Phonation during Connected Speech in Adductor Spasmodic Dysphonia

Michael P. Cannito*, Eugene H. Buder1, Lesya B. Chorna1 and Richard Dressler2

1School of Communication Sciences and Disorders, University of Memphis, Memphis, TN, USA
2Department of Communication Disorders, Western Kentucky University, Bowling Green, KY, USA

Abstract

Objectives: This study examined acoustic measures related to voice production in connected speech of patients with adductor spasmodic dysphonia (ADSD) before and after botulinum toxin (BT) treatment in comparison with speech of non-dysphonic healthy control (NHC) speakers. Correlations between acoustic measures and perceptual scaling judgments of overall voice quality were also examined.

Methods: Ten patients with moderate-to-severe ADSD were audio-recorded prior to and following BT injection, as were age and gender matched NHCs. Signal processing algorithms were employed to extract fundamental frequency (f0) and intensity (dB) of speakers’ digitized oral readings. Control procedures minimized the influence of severely aperiodic phonation and obstructed consonant production on the analysis of modal f0 and related variables such as signal-to-noise ratio (SNR). Measures of severe aperiodicity, with f0 < 80 Hz, were compared with modal f0 data. Perceptual judgments of voice quality were obtained from expert voice clinicians under rigorously controlled listening conditions.

Results: Acoustic measures of number of low frequency f0 events, coefficient of variation of modal f0, and SNR demonstrated statistically significant differences for ADSD speech before and after treatment, and differentiated between ADSD and NHC speakers. Mean modal f0 did not differentiate among speaker conditions. All other measures demonstrated untreated ADSD speech from that of NHCs; however, only selected measures demonstrated differences between NHC and ADSD speakers following BT injection. Coefficient of variation of modal f0 and SNR were moderately correlated with expert listener judgments of voice quality.

Conclusion: Acoustic measures of connected speech derived from carefully edited modal f0 tracks and intensity contours were effective for characterizing ADSD speech before and after treatment, and for differentiating it from that of NHCs. Acoustic measures were highly reliable and significantly related to voice quality scaling by expert listeners.

Keywords: Adductor spasmodic dysphonia; Phonatory acoustic measures; Botulinum toxin

Introduction

Adductor spasmodic dysphonia (ADSD) is an unusual voice disorder resulting from laryngeal dystonia and for which there is a growing body of evidence for neurogenesis of the condition [1,2]. ADSD voice is characterized by the presence of intermittent aperiodicity, particularly of a “strained-strangled” quality, along with frequent voice breaks and abrupt shifts of fundamental frequency (f0). These are the result of unpredictable involuntary spasms of the vocal folds in the direction of glottal closure during the production of voiced speech. A variety of acoustic measures have been employed to study ADSD, prior to and/or following medical treatments, such as botulinum toxin (BT) injection of the vocal fold(s) [3].

The majority of these studies have focused on the acoustical analysis of sustained vowel phonation [4-11]. Yet, phonation in connected speech must be the ultimate criterion by which we judge the functional communicative significance of ADSD as well as its response to treatment [12]. Moreover, in the case of ADSD, there are clinically significant differences in severity between phonation in the context of sustained vowels as compared to connected speech [12,13]. Acoustic analyses of connected speech in ADSD have been uncommon. Early studies of connected speech were of limited scope. For example, Ford, Bless and Patel [14] demonstrated decreased oral reading time from pre-to-post treatment, whereas Whurr et al. [15] observed reduced standard deviation of f0 following treatment. A major reason for such limited analyses is that periodicity-based signal processing algorithms used to measure jitter, shimmer, or signal-to-noise ratio (SNR), may be inappropriate for the analysis and characterization of connected speech for severely disordered voices [12,16]. This is based on the fact that the algorithms used to detect f0 for pitch extraction are designed to identify relatively regular periodicity in the ongoing waveform and are prone to break down in the presence of extremely erratic or aperiodic signals [17]. Of course, such aperiodic signals are commonplace in ADSD voice. In addition, extraction of f0 in connected speech is generally hampered by the destabilizing influence of obstructed consonant production and associated aperiodic noise concurrent with phonation.

In an effort to better understand and quantify voice produced in connected speech by speakers with ADSD, some researchers have employed a strategy of counting the frequency of occurrence of well-specified acoustic events (i.e., aperiodicity, f0 shifts, and voice breaks) that are visible in the acoustic waveform and are related to vocal fold...
dysfunction in ADSD, and calculating their percentage of cumulative duration relative to utterance duration [12,18]. Other researchers have used the method of signal typing [17], which employs qualitative judgments of the time varying spectrum such as absence versus presence of aperiodicity, \( f_0 \) bifurcations, or voice breaks [19] and determining their percentage of occurrence relative to overall voiced segment duration in the connected speech of ADSD.

Although these measures have demonstrated some utility for clinical research, one limitation of such methods for analyzing connected speech in severely disordered voices is that, even though they can be counted and their duration measured, they remain underlying categorical judgments, albeit based on inspection of physical acoustic data. As such they are discontinuous events of underlying nominal scale, and as such may be less sensitive to small differences (as a result of treatment, for example) than are continuous ratio level measurements of frequency and intensity varying over time within the voice signal. The purpose of the present study is to demonstrate the effectiveness of some alternative approaches to extraction and measurement of \( f_0 \) and associated features of amplitude and SNR in connected speech. These are both continuous and ratio in scale, and were obtained from an archival data base of recordings of speakers with ADSD pre/post BT treatment and of non-dysphonic healthy control (NHC) speakers [20]. Such measures may prove useful toward developing a more complete profile of voice characteristics in ADSD before and after treatment. Specifically, it is hypothesized that the judicious application of off-line editing of the \( f_0 \) tracks taken from continuously voiced segments of ADSD speech with \( f_0 \) occurring above a threshold of 80 Hz, will yield measures that are reliable and sensitive to differences between connected speech of speakers with ADSD before and after BT treatment, and between ADSD and NHC speakers. The \( f_0 \) data have been carefully hand-edited to remove voice episodes that are extremely aperiodic, to correct for subharmonics, and to eliminate destabilizing effects of consonant articulation. These so called “high \( f_0 \) track” measures also were compared with measures based on the categorical occurrence of phonatory activity below 80 Hz (i.e., strained-strangled dysphonia and associated voice breaks) that are not measured as continuous ratio level data. In addition, the relationship of the acoustic measures to carefully controlled perceptual scaling judgments of voice quality made by expert voice clinicians was examined.

Materials and Methods

Participants and speech samples

Connected speech samples from 10 female speakers diagnosed with ADSD for at least four years prior to the study were selected from a larger archival database of recordings from 42 consecutive cases with ADSD who underwent initial BT injection [20]. Their clinical characteristics are provided in Table 1. An otolaryngologist and a speech-language pathologist familiar with the disorder, which has been shown to be reliable in previous research [12,22]. All speakers were judged to be free of laryngeal tremor and did not exhibit movement disorders elsewhere in the body. ADSD speakers underwent EMG guided transectioneaneous injection of BT (type A) into either the left (\( n = 8 \)) or both (\( n = 2 \)) vocal folds administered by a licensed otolaryngologist who was familiar with the procedure, with dosages varying from 2.5 to 15 Units.

Connected speech samples consisted of an oral reading of the first paragraph of the Rainbow Passage [21]. The selected recordings represented a midrange of pre-injection severity of ADSD, based on a five-point ordinal rating scale determined by consensus of two speech pathologists familiar with the disorder, which has been shown to be reliable in previous research [12,22]. This yielded ten pre-injection recordings in the moderate to severe ranges as well as ten post-injection recordings produced by the same speakers. A midrange of dysphonic severity was targeted for study because, given the intermittent nature of ADSD, mild cases would tend to skew the data toward non-significance.

<table>
<thead>
<tr>
<th>Subject No.</th>
<th>Age</th>
<th>TPI</th>
<th>Units</th>
<th>Injection</th>
<th>TPO</th>
<th>Pre Severity</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>46</td>
<td>34</td>
<td>15</td>
<td>Unilat</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>16</td>
<td>38</td>
<td>38</td>
<td>15</td>
<td>Unilat</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>18</td>
<td>63</td>
<td>31</td>
<td>15</td>
<td>Unilat</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>25</td>
<td>56</td>
<td>31</td>
<td>15</td>
<td>Unilat</td>
<td>13</td>
<td>2</td>
</tr>
<tr>
<td>33</td>
<td>60</td>
<td>37</td>
<td>15</td>
<td>Unilat</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>34</td>
<td>31</td>
<td>45</td>
<td>15</td>
<td>Unilat</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>39</td>
<td>52</td>
<td>35</td>
<td>5</td>
<td>Unilat</td>
<td>24</td>
<td>2</td>
</tr>
<tr>
<td>42</td>
<td>67</td>
<td>51</td>
<td>2.5</td>
<td>Bilat</td>
<td>13</td>
<td>3</td>
</tr>
<tr>
<td>45</td>
<td>38</td>
<td>13</td>
<td>2.5</td>
<td>Unilat</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>50</td>
<td>46</td>
<td>43</td>
<td>2.5</td>
<td>Bilat</td>
<td>9</td>
<td>3</td>
</tr>
</tbody>
</table>

Severity: 0 - normal; 1 – mild; 2 – moderate; 3 – severe; 4 – profound. TPO: time post onset of SD, years; TPI: time post injection, days.

Table 1: Clinical characteristics of 10 female speakers with ADSD
It is important to acknowledge that these archival recordings did not conform to contemporary standards now generally preferred for absolute quantification of most voice quality measures such as perturbations and SNRs. In particular, cassette tapes are significantly inferior to direct digitizing [24] but the effects are absolute and quantifiable as offsets from ideal [25] and so many researchers have found work with such media to be acceptable for certain purposes. Here, the focus on comparing pre- and post-treatment characteristics, with all participants recorded using the same technology is the quantification of interest, not necessarily a claim that exact low levels of perturbation are reflective of exact vocal fold cycle-to-cycle variations per se. Additionally, higher sampling rates are preferable for absolute quantification of departures from pure periodicity [26] as high as 50 kHz or more, but for the algorithms to be utilized in this work it has been demonstrated that rates on the order of 20 kHz are acceptable [27] especially for the current design which focuses on relative levels of the measures.

**Procedures**

The first, fourth and sixth sentences of the Rainbow Passage were excerpted from each recording: “When the sunlight strikes raindrops in the air, they act like a prism and form a rainbow”, “These take the shape of a long, round arch with its path high above and its two ends apparently beyond the horizon”, “When a man looks for something beyond his reach, his friends say he is looking for the pot of gold at the end of the rainbow”. These yielded sixty utterances for acoustic analysis. Measures were based on the following features: total sentence duration (TSD); f0 contour; intensity contour (for which f0 was present above 80 Hz); and SNR.

Acoustic analysis involved two stages of digital signal processing. The first stage was extraction of basic data: total sentence duration (TSD) as well as extraction of fundamental frequency contours, intensity contours and signal-to-noise ratios across the utterance. The second stage was derivation of acoustic variables from extracted data.

TSD was calculated using Milenkovic’s CSpeech software [28]. Corresponding waveforms and spectrograms were displayed on the monitor screen, and both visual inspection and auditory playback were used to guide cursor placement at the onset and offset of acoustic energy for the first and last phonetic segments of the sentence, and time between cursors was calculated. Inter-analyst reliability was established for TSD. Ten percent of all utterances were re-measured by two researchers. Correlation coefficient for inter-analyst reliability for utterance duration was $r = 0.98$. Means and standard deviations of TSD for three speaker conditions averaged across three sentences are shown in Table 2. Analysis of variance (ANOVA) determined that the TSD differed significantly ($p<0.05$) between speaker conditions as well as between sentences, but there were no significant interactions. No significant difference was observed within speakers with ADSD from pre- to post-BT injection; however, post hoc tests demonstrated that ADSD subjects differed from normal controls before and after BT injection ($p<0.05$). These analyses suggest that while duration was sufficiently stable in ADSD to allow for meaningful comparisons of fundamental frequency and intensity derived measures before and after treatment, TSD differences between ADSD and controls indicated a need to convert temporal measures of phonation time into percentages for the subsequent analysis.

**Fundamental frequency analyses**: Automated f0 contour, intensity contour (Adj dB), and SNR, which utilized algorithms from the CSpeech software [28], were calculated using CSpeech (Milenkovic,1997) for each utterance. Only modal voiced segments were desirable for analysis, therefore obstruent phonemic productions were excluded.
from analysis with the exception of taps and flaps. Cursors were placed by hand to select the segments of interest, as guided by spectrographic inspection and listening. Modal phonation data were submitted to periodicity-based analyses of fundamental frequency, intensity and signal-to-noise ratio. The pitch determination algorithm implemented in CSpeech for f0 extraction was of the center-clipped autocorrelation type, with parameters available for controlling several aspects of the procedure, including f0 range. The periodicity determination for SNR extraction was more extensive, using the core principles documented in Milenkovic [29] and with improvements documented in Milenkovic [30]. Fundamental frequency contours were saved as ASCII files for further analysis. Frequency contours were displayed in CSpeech to make possible their visual inspection for hand editing. The purpose of hand editing was to analyze individual glottal pulses via cursor placement. Any aperiodic voiced segments (e.g. marked breathiness or roughness) were excluded from f0 analysis. A frequency floor of 80 Hz was employed to restrict the analysis to modal voice and exclude pulse register phonation and other types of voice breaks from the analysis. Any voiced segments of 30 or more milliseconds in duration dropping below the floor were also excluded from automated pitch extraction procedure. These segments were marked and labeled ‘sde’ (spasmodic dysphonie episode) for subsequent inspection.

The autocorrelation algorithm from CSpeech was used for the extraction of the f0 contour. Algorithm parameters were optimally set to yield consistent measures from both men and women with ADSD. For all participants, the f0 frame shifts were set to the default value of 5 ms and the “downsampling factor” was set to the recommended 3; these settings imposed consistent smoothness and general precision standards across all the samples. Other parameters included a window length of 15 ms and an f0 range of 80-300 which was appropriate for women. The f0 contour was then examined for accuracy and hand edited when the algorithm failed by either eliminating spurious values or by assigning correct values by inspection for glottal pulses in the waveform and/or spectrogram. All excluded segments appeared as a flat baseline in the tracks at the “zero” level (actually a non-observation). In some signals, presence of subharmonics would draw the fundamental frequency contour below 80 Hz. In such cases, hand-editing inspection was performed to assign the fundamental frequency contour values to the level of the true first harmonic level, not the subharmonic one.

Reliability of hand editing was evaluated by having two different researchers, who were trained in the procedures, compare f0 contour from 10% of the overall data. This included three sentences from three normal, three pre, and three post-BT injection recordings yielding a total of 27 f0 contours. Within each file, the researchers identified data points in need of editing. Thirty-five percent of each f0 contour were hand edited on average (S.D.=5%). Percentage of hand edited f0 points was similar across all speaking conditions. Point-for-point agreements were tabulated for each of 27 utterances. Average percent inter-researcher agreement for the edited f0 contours was 89.33%. Overall mean absolute difference in f0 between the researchers was 6.85Hz, with S.D. of 15.53Hz.

Modal phonation time was defined as time when edited fundamental frequency contour was present (between 80-300 Hz). Average modal phonation time for each speaking condition is reported in Table 2. ADSD subjects have longer phonation time, but also have significantly longer TSD (see Table 2). To compensate for the duration differences, percent phonation time was calculated and used for statistical analysis. Percent modal phonation time, based on the ratio of modal phonation time and total sentence duration, is defined as the following:

\[
\% f0_{\text{mod}} = \frac{Tf0_{\text{mod}}}{TSD} \times 100\%
\]

Where TSD denotes total utterance duration time for a sentence, Tf0_{\text{mod}} is modal phonation time for a sentence.

Very low fundamental frequency segments (e.g. glottal pulsing with fundamental frequency below 80 Hz) were marked as “low fundamental frequency activity” or “low tracks” meaning that fundamental frequency contour existed in those places but was not tracked because of 80 Hz lower limit setting in the algorithm. The low tracks were of interest because they corresponded to the perceptual sign of “strained-strangled” voice quality that typifies ADSD and is frequently associated with voice breaks. These episodes were flagged with a fundamental frequency value of zero. They were inspected using a combination of waveform and spectrogram displays as well as audio playback to verify the presence of pulse phonation. Durations of low tracks were measured, and number of low track events was counted for each sentence of 30 samples. Mean durations of low track events were calculated by dividing total low track durations by number of low track events. Percent low track phonation time was calculated as a ratio of low track duration and TSD as the following:

\[
\% f0_{\text{low}} = \frac{Tf0_{\text{low}}}{TSD} \times 100\%
\]

Intensity contours (Adj dB) were extracted from the waveforms using a RMS algorithm from the CSpeech software [28] and were 20 ms window length, and the intensity contours were saved as ASCII files. These were also displayed for visual inspection. To provide an intensity reference, noise floor segments were extracted from the corresponding waveform files. Noise floor was defined as approximately 200 ms section of the wave file that was absent of audible respiration and speech/non-speech noise. If a single 200 ms section could not be found then two 100 ms sections were obtained. Intensity contours from extracted connected speech were edited in ASCII files prior to averaging. Only intensity values corresponding to the time frames at which fundamental frequency existed were used to calculate average signal intensity. Average intensity of the noise floor (RMS) was calculated

<table>
<thead>
<tr>
<th>Condition</th>
<th>NHC</th>
<th>ADSD PRE-BT</th>
<th>ADSD POST-BT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>6602.7</td>
<td>8997.3</td>
<td>7547.8</td>
</tr>
<tr>
<td>S.D.</td>
<td>1348.7</td>
<td>2960.5</td>
<td>2027.5</td>
</tr>
<tr>
<td>High Track Phonation Time</td>
<td>4059.0</td>
<td>4430.2</td>
<td>4345.3</td>
</tr>
<tr>
<td>S.D.</td>
<td>669.6</td>
<td>820.4</td>
<td>881.4</td>
</tr>
<tr>
<td>Low Track Phonation Time</td>
<td>25.13</td>
<td>382.43</td>
<td>124.19</td>
</tr>
<tr>
<td>S.D.</td>
<td>56.52</td>
<td>384.18</td>
<td>110.46</td>
</tr>
</tbody>
</table>

Table 2: Means and standard deviations of the speaking time and phonation times across three sentences.
from a file with 200ms of silence. Intensity values were converted to decibels using the following formula:

\[
\text{Adjusted dB} = 20 \log_{10} \frac{\text{RMS}_n}{\text{RMS}_0} + 25
\]

An arbitrary number of 25 dB in this formula was used to adjust for the ambient noise.

**Signal-to-Noise (SNR) analyses:** SNR values were calculated for those segments of the f0 contour which met the following criteria: 1) segments were at least 100 ms in length; 2) segments tracked with the CSpeech Newjit algorithm (Analysis-Voice-Newjit) had average f0 values approximately the same as in the hand inspected fundamental frequency contours at the same time points. The latter criterion is especially important for guaranteeing that the periodicity-based SNR was calculated against the same periodicity basis as obtained by validated f0 ASCII files with the results from Newjit were saved.

A careful two-step screening procedure was applied before average SNR values for each sentence were calculated. The first step consisted in retaining only the SNR values that correspond to correct f0 contour values from hand inspected pitch traces (with tolerance level ±10Hz). The second step was to examine the values which had been already screened for correct f0 SNR (in step 1) and retain only those for which the CSpeech Newjit algorithm reported an "error" value of 2 or less [31]. An "error" term is a flag indicating uncertainty in fundamental frequency period estimation. Fundamental frequency estimation in the CSpeech Newjit algorithm is based primarily on a cross-correlation algorithm. The algorithm estimates current f0 based on the previous values of f0. When a current value is difficult to determine for any reason, the program reports f0 flagged with the "error" term: the bigger the "error" number, the greater the possibility of wrongly identified f0. For comparison purposes, the unscreened SNR values were also recorded. The overall numbers of 100ms segments (as reported by the CSpeech algorithm), numbers of frames rejected due to screening procedure, and numbers of frames included for average SNR calculation were ASCII files with the results from Newjit were saved.

The overall numbers of 100ms segments (as reported by the CSpeech algorithm), numbers of frames rejected due to screening procedure, and numbers of frames included for average SNR calculation were calculated. The results are shown in Table 3, wherein after the screening procedures, 60% of frames were retained for NHC subjects, 32% for ADSD in Pre-BT condition, and 36% in ADSD in Post-BT condition.

**Derived variables used for statistical analysis:** The following measures were derived from the TSD, f0, and intensity data, and served as dependent variables for further statistical analyses. Each measure was calculated for three sentences in three speaking conditions (ADSD, before and after BT injection, and NHC).

**Low Track Analyses**

1. Total low track duration (in ms);
2. Number low track events;
3. Mean low track duration (in ms);
4. Percent low track phonation time.

**High Track Analyses**

1. Mean fundamental frequency (Average f0 in Hz);
2. Coefficient of variation of f0 (COV f0);
3. Mean intensity (Adjusted dB);
4. Coefficient of variation of intensity (COV Adjusted dB);
5. Average SNR (screened for correct f0 with number "errors" ≤ 2);
6. Percent modal phonation time.

Inter-analyst reliability was established for derived high track measures for 27 utterances. Pearson correlations (r) were 0.99 for Average f0, 0.95 for COV f0, 0.99 for Adj dB, 0.94 for COV Adj dB, 0.99 for Average SNR, and 0.92 for Percent Modal Phonation time. High correlation coefficients suggest that the measures can be reliably reproduced. Paired sampled t-tests revealed no significant differences between the measurements taken by two independent researchers.

**Perceptual Scaling of Voice Quality:** Archival data also were employed to determine the relationship between the current acoustic measures and perceived severity of dysphonia. In the original study [20], from which the current participants were sampled, six expert voice clinicians served as listeners to provide computer implemented visual analog scaling judgments of overall voice quality under carefully acoustically controlled listening conditions. Auditory signals (the recorded utterances) were randomly presented to individual listeners via loud speaker in a sound treated booth at 55 dB SPL. Listener judgments were highly reliable (intra-listener r=0.96; inter-listener ICC = 0.98) and were obtained for speakers with ADSD pre-BT, post-BT, and for NHC speakers. In the present study, the mean voice quality scaling data, averaged across six listeners, for each of the current of speakers was correlated with the acoustic variables using Pearson product moment correlation coefficients at alpha level = 0.01.

**Results**

**Low track analysis**

Descriptive statistics for low track data are presented in Table 4. A statistically significant (p<0.001) two-way multivariate analysis of variance (MANOVA) with repeated measures was computed for the low track data, followed by univariate ANOVAs for the individual low track variables (speaker condition by sentences). Results for the main effect of speaker conditions are provided in Table 5. All low track ANOVAs yielded statistical significance (p<0.001) for the main effects of speaker conditions. There were no significant main effects or interactions involving sentences in these analyses.

Post hoc means comparisons were calculated using Tukey's HSD statistic (p= 0.05). It was demonstrated that ADSD speakers' low track values decreased significantly from Pre-to-post BT injection for Total Low Track Duration, Number of Low Track Events, and Percent Low Track Duration. ADSD speakers exhibited significantly greater values than NHC speakers prior to BT treatment for Total

<table>
<thead>
<tr>
<th>Variable</th>
<th>Unscreened SNR</th>
<th>Screened SNR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NC</td>
<td>PRE</td>
</tr>
<tr>
<td>Mean SNR</td>
<td>23.38</td>
<td>15.95</td>
</tr>
<tr>
<td>S.D. SNR</td>
<td>1.66</td>
<td>2.56</td>
</tr>
<tr>
<td>Number of frames used to calculate SNR</td>
<td>31.36</td>
<td>32.07</td>
</tr>
<tr>
<td>% of points retained</td>
<td>n/a</td>
<td>n/a</td>
</tr>
</tbody>
</table>

Table 3: Screened and unscreened SNR values and number of frames used to calculate SNR across three speaker conditions.
Low Track Duration, Number of Low Track Events, Mean Low Track Duration and Percent Low Track Duration. Following BT treatment, ADSD speakers continued to exhibit significantly greater values than NHC speakers for Number of Low Track Events and Mean Low Track Duration.

### High track analysis

Descriptive statistics for high track data are presented in Table 6. A statistically significant (p < 0.001) two-way MANOVA with repeated measures was computed for the high track data, followed by univariate ANOVAs for the individual high track measures (speaker condition by sentences). There were no significant main effects or interactions involving sentences in these analyses. Statistical results are summarized in Table 7. No significant main effect of speaking conditions was observed for the average $f_0$ (p > 0.10). Statistically significant main effects of speaking conditions (p < 0.001) were obtained for the remaining high track variables of COV $f_0$, Adj dB, COV Adj dB, SNR, and percent $f_0$ time.

Post hoc testing (Tukey’s HSD; p < 0.05) demonstrated that ADSD speakers’ COV $f_0$ decreased significantly from pre-to-post BT treatment, while SNR and Percent $f_0$ time increased significantly from pre-to-post BT. Prior to BT treatment, speakers with ADSD exhibited significantly greater values than NHC speakers for COV $f_0$ and COV Adj dB, but significantly smaller values than NHC speakers for Adj dB and SNR. Following BT treatment, speakers with ADSD continued to exhibit greater values than NHC speakers for COV $f_0$ and COV Adj dB, in spite of treatment-related improvement for these variables. Also following BT treatment, speakers with ADSD continued to exhibit significantly smaller values than NHC speakers for Adj dB and SNR, in spite of treatment related improvement in these variables.

### Correlation of acoustic variables with perceived voice quality

Correlation of acoustic measures with perceived voice quality demonstrated significant relationships (p < 0.01) for a number of acoustic variables. High track variables that were correlated with voice quality included COV $f_0$ (r = -0.689), Adj dB (r = 0.693), COV Adj dB (r = -0.478), Screened SNR (r = 0.782), and Percent $f_0$ time (r = 0.525). Positive r values indicated that the greater the acoustic measure, the better the voice quality, while negative values indicated the greater the acoustic measure, the poorer the voice quality. No significant correlation was observed for Mean $f_0$. None of the low track variables were correlated significantly (p > 0.05) with perceived voice quality.

### Discussion

The purpose of this study was to demonstrate the efficacy of acoustic measures of $f_0$ and its associated features (e.g., SNR), extracted from connected speech in speakers with ADSD and NHC speakers, as well as to compare the clinical utility of these ratio level measures with acoustic measures derived from categorical judgments of dysphonia occurrence. The relationships of both types of measures to expert clinicians’ perceptual scaling judgments of voice quality also were examined. Results demonstrated that several of the acoustic measures were useful for differentiating ADSD speech from that of NHC speakers before or after treatment, and for demonstrating pre-to-post treatment change in the speech of individuals with moderate-to-severe ADSD. The three most sensitive measures, those which significantly differentiated all speaker conditions from each other, were Number of Low Track Events, COV $f_0$ and Screened SNR. Other measures that demonstrated treatment-related change included Total Low Track Duration and Percent Low Track Duration, as well as the high track variable of Percent $f_0$ time; however, each of these measures proved insensitive to differences between the speech of patients with ADSD post-BT treatment and that of the NHC speakers. With the exception of Mean $f_0$, all other measures differentiated ADSD speech prior to treatment from that of the NHCs. Mean Low track duration, Adj dB and COV Adj dB also differentiated ADSD speech post-treatment from that of the NHCs, but did not differ from pre-to-post treatment ADSD.

---

**Table 4**: Means (standard deviations) for low track variables across three speaker conditions*

<table>
<thead>
<tr>
<th>Conditions</th>
<th>Total low track duration</th>
<th>Number of low track events</th>
<th>Mean low track duration</th>
<th>Percent low track duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal Controls</td>
<td>25.13 (58.52)</td>
<td>0.17 (0.38)</td>
<td>25.13 (58.52)</td>
<td>0.35 (0.81)</td>
</tr>
<tr>
<td>ADSD Pre BT</td>
<td>382.43 (384.18)</td>
<td>4.20 (3.67)</td>
<td>84.17 (72.19)</td>
<td>4.38 (4.88)</td>
</tr>
<tr>
<td>ADSD Post BT</td>
<td>124.19 (110.46)</td>
<td>1.63 (1.22)</td>
<td>66.10 (45.92)</td>
<td>1.79 (1.58)</td>
</tr>
</tbody>
</table>

* Wilk’s Lambda = 0.509

**Table 5**: Statistical results for main effect of speaker conditions on low track variables*

<table>
<thead>
<tr>
<th>Conditions</th>
<th>Mean $f_0$ (4.72)</th>
<th>COV $f_0$ (0.04)</th>
<th>Adj dB (3.18)</th>
<th>COV Adj dB (0.01)</th>
<th>Screened SNR (1.62)</th>
<th>Percent $f_0$ time (5.86)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal Controls</td>
<td>182.94 (169.60)</td>
<td>0.17 (0.06)</td>
<td>62.15 (47.65)</td>
<td>0.09 (0.09)</td>
<td>24.71 (18.59)</td>
<td>25.13 (58.52)</td>
</tr>
<tr>
<td>ADSD Pre BT</td>
<td>169.60 (25.31)</td>
<td>0.26 (0.05)</td>
<td>47.65 (5.76)</td>
<td>0.09 (0.02)</td>
<td>16.18 (3.32)</td>
<td>25.13 (58.52)</td>
</tr>
<tr>
<td>ADSD Post BT</td>
<td>172.90 (25.11)</td>
<td>0.20 (0.06)</td>
<td>48.91 (3.62)</td>
<td>0.08 (0.02)</td>
<td>18.06 (4.30)</td>
<td>172.90 (25.11)</td>
</tr>
</tbody>
</table>

**Table 6**: Means (standard deviations) for high track variables across three speaker conditions.

---

The performance of screened SNR affirmed the efforts taken to apply this dysphonia measure to connected speech. By restricting SNR calculation to intervals over which the \( f_0 \) periodicity basis was below 80 Hz and corroborated by hand inspection of pitch traces, and by retaining for analysis only those for which the algorithm reported minimal \( \text{Adj dB} \) errors, the procedure guaranteed that problems with periodicity detection were minimized and allowed valid application of the algorithm to connected speech. The absolute values may have been elevated somewhat by the original use of cassette recordings and only a moderately high sampling rate (20 kHz). On the other hand, the application procedures helped to assure that more egregious spurious noise measures produced by algorithm failure, incorrect \( f_0 \) determination, and inclusion of highly variable periods generally found during low \( f_0 \) were maximally reduced. At least one report has called into question the validity of periodicity detection in the CSpeech/TF32 to produce SNR and perturbation measures generally found during low \( f_0 \) which did not differentiate between any speaker conditions. It is unsurprising that, having removed all \( f_0 \) values less than 80 Hz, the mean \( f_0 \) was essentially normal and similar across the speaking conditions. This finding substantiates the fact that ADSD is an intermittent disorder and that most patients demonstrate episodes of relatively normal voice when speaking [3]. The finding that Number of Low Track Events was the most sensitive of the categorically-based, low track variables is consistent with the diagnostic criterion of occurrence of adductor vocal spasms, in the presence of intermittent normal voicing, as the primary diagnostic feature of the disorder [3]. It is interesting that the \( f_0 \)-derived high track variables of Screened SNR and COV \( f_0 \) were at least as sensitive as Number of Low Track events to differences among the three speaking conditions. These findings are in good agreement with those of Sapienza et al. [12] who observed the occurrence of aperiodicity and pitch shifts to be important acoustic markers for ADSD in comparison to NHC speech, as well as for demonstrating pre-to-post treatment improvement in ADSD. The advantage of the present measures, however, is that they provide direct quantification of the actual magnitude of \( f_0 \) variability and aperiodicity that is present in the waveforms. In addition, SNR and COV \( f_0 \) yielded moderately strong correlations with expert listeners' scaling of perceived voice quality that were not exhibited by the categorically derived low track variables. This finding suggests that although vocal spasms (low track episodes) may be pathognomonic for ADSD, the severity of dysphonia was based on the amount of aperiodicity and \( f_0 \) variability.

The role of signal intensity (\( \text{Adj dB} \)) was also of particular interest in the current data set. NHC speakers were notably louder than ADSD speakers before and after treatment, and intensity did not improve post-treatment. Increased glottal resistance accounts for reduced vocal intensity in untreated ADSD; however, glottal resistance improves following BT injection [33]. It is likely that the weakening effect of BT on vocal fold muscle activation may have contributed to reduced intensity observed post-treatment since the post treatment voice has been shown to be somewhat breathy during the post injection period included in this study [20].

In conclusion, the novel approach to dealing with \( f_0 \) data in connected speech employed in the present study appears promising for application to describing voice disorders and monitoring effect of treatment. Due to inherent limitations of the archival data, in future studies it will be desirable to replicate these findings using direct digitization of speech signals with a minimum sampling rate of 44.1 kHz. It should also be recognized that the hand editing process for retention of \( f_0 \) data was highly labor intensive and time consuming, and may therefore be of limited clinical applicability. The advent of cepstral analysis applied to connected speech [34,35] may prove to be a more efficient method for monitoring aperiodicity in ADSD in future studies, although these approaches necessarily incorporate all speech segments and don’t focus exclusively on phonation in the manner that the present study did. Alternatively, the use of harmonic amplitude differences (e.g., H1-H2) also offers a viable strategy for quantifying voice in connected speech of ADSD [36].

Acknowledgement

The authors are indebted to Gayle E. Woodson, M.D. and Thomas Murry, Ph.D. for providing the patient recordings for this study. We also gratefully acknowledge Brenda Bender, Ph.D. for assisting with the reliability analyses.

References


