EURASIP Journal on Applied Signal Processing 2005:9, 1400–1409

Disordered Speech Assessment Using Automatic Methods Based on Quantitative Measures

Lingyun Gu
Computational NeuroEngineering Laboratory, Department of Electrical & Computer Engineering, University of Florida, Gainesville, FL 32611-6200, USA
Email: lygu@cnel.ufl.edu

John G. Harris
Computational NeuroEngineering Laboratory, Department of Electrical & Computer Engineering, University of Florida, Gainesville, FL 32611-6200, USA
Email: harris@cnel.ufl.edu

Rahul Shrivastav
Department of Communication Sciences & Disorders, University of Florida, Gainesville, FL 32611, USA
Email: rahul@csd.ufl.edu

Christine Sapienza
Department of Communication Sciences & Disorders, University of Florida, Gainesville, FL 32611, USA
Email: sapienza@csd.ufl.edu

Received 2 November 2003; Revised 6 August 2004

Speech quality assessment methods are necessary for evaluating and documenting treatment outcomes of patients suffering from degraded speech due to Parkinson's disease, stroke, or other disease processes. Subjective methods of speech quality assessment are more accurate and more robust than objective methods but are time-consuming and costly. We propose a novel objective measure of speech quality assessment that builds on traditional speech processing techniques such as dynamic time warping (DTW) and the Itakura-Saito (IS) distortion measure. Initial results show that our objective measure correlates well with the more expensive subjective methods.

Keywords and phrases: objective speech quality measures, subjective speech quality measures, pathology, anthropomorphic.

1. INTRODUCTION

The accurate assessment of speech quality is a major research problem that has attracted attention in the field of speech communications for many years. The two major classes of methods employed in the assessment of speech quality are subjective and objective speech quality measures. Subjective quality measures are more accurate and robust since they are given by professional personnel who have received special assessment training, but they are necessarily time consuming and costly. On the contrary, objective quality measures, inspired by speech signal processing techniques, provide an efficient, economical alternative to subjective measures. Although it is not suggested to use objective quality measures to completely replace subjective measures, objective quality measures do show the strong ability to predict subjective quality measures and the results do correlate very well with those produced by subjective quality measures [1]. Traditionally, objective measures have been used to evaluate speech after decoding and in the presence of noise. Currently, some pioneers have already developed some system protocols or algorithms to apply objective speech quality assessment into disordered speech analysis.

Any meaningful quality assessment should be consistent with human responses and perception. Therefore, subjective measures naturally became the first choice to evaluate speech quality. Performance methods using subjective measures are based on a group of listeners' opinion of the quality of an utterance. Subjective measures usually focus on speech intelligibility and the overall quality. Subjective measures can also be broadly grouped into two categories: utilitarian and analytic. Utilitarian methods have three goals: (1) they should be reasonably efficient in test administration and data analysis; (2) they evaluate speech quality on a unidimensional scale;

(3) they must be reliable and robust in their test method. The key aspect of utilitarian approaches is that the results are summarized by a single number. On the other hand, analytic methods try to identify the underlying psychological components that determine perceived quality, and to discover the acoustic correlates of these components. Therefore the results from analytic methods are summarized on a multidimensional scale [1].

The modified rhyme test (MRT) by House and the diagnostic rhyme test (DRT) by Voiers are both intelligibility measures. The mean opinion score (MOS) test and the diagnostic acceptability measure (DAM) are overall quality measures, even though MOS is also commonly categorized as utilitarian and DAM is classified as analytic. It is understandable that subjective quality measures are the preferable means of quality assessment but subjective measures do have several major drawbacks: (1) subjective measures require significant time and personnel resources, making it difficult to evaluate the range of potential speech/voice distortion; (2) subjective measures do not work very well when the tested speech database is large [2]; (3) some rating score protocols are not suitable for measurement of speech/voice [3]; (4) some literature suggests that listeners cannot agree on specific speech/voice ratings [4].

Compared with the subjective measures mentioned above, objective measures have several outstanding advantages: (1) they are less expensive to administer, saving money, time, and human resources; (2) they produce more consistent results and are not affected by human error; (3) most importantly, the form of the objective measure itself can give valuable insight into the nature of the human speech perception process, helping researchers understand the speech production mechanism more deeply [1]. Generally speaking, objective speech quality measures are usually evaluated in the time, spectral, or cepstral domains.

This paper is organized as follows. In Section 2, disordered speech background will be introduced. Then, in Section 3, the DTW method is discussed. Specific speech features for disordered speech will be proposed in Section 4. Section 5 deals with one subjective measure. All experimental results are discussed in Section 6. Finally, conclusions are drawn in Section 7.

2. DISORDERED SPEECH BACKGROUND

Usually, patients with Parkinson's disease or people who have suffered a stroke have difficulty producing clear speech, resulting in a loss of intelligibility. Hence, it is important to develop a means to help them produce more clear speech or develop algorithms to automatically clarify their unclear speech. These efforts require an efficient method to evaluate disordered speech as the first step.

Attempts to develop algorithms to evaluate disordered speech require us to understand how disordered speech is produced, the factors that affect disordered speech, and the explicit phenomena related to these factors. The term "dysarthria" is used to describe changes in speech production

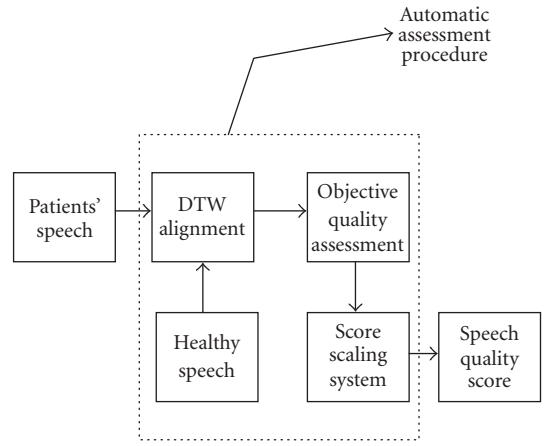

Figure 1: Objective patients' speech quality assessment block diagram.

characterized by an impairment in one or more of the systems involved in speech [5]. The three major systems involved in speech production are respiration, voice production, and articulation. Voice is produced by the larynx and the oral structures articulate to modify the sound source produced by the larynx. The dysarthria associated with Parkinson's disease is referred to as a hypokinetic dysarthria [6, 7]. Common symptoms of hypokinetic dysarthria include reduced loudness of speech and/or monoloudness (lack of loudness variation) and reduced speaking rate with intermittent rapid bursts of speech. For instance, speakers may show a slow rate of speech, but particular words or phrases within that utterance may be produced with a rapid rate. The oral structures such as the tongue and lips are "rigid," resulting in a reduced range of movement. This effectively dampens the speech signal and distorts the accuracy of the sound (consonant or vowel) production. There may be some instances of hypernasality as the condition worsens resulting from an inadequate velar closure. This may also result in the dampening of the sound produced. Voice quality in these patients is often described as hoarse or harsh.

In this paper, we test several well-known speech processing parameters that can quantify the severity of disordered speech. These are the Itakura-Saito (IS) measure, the log-likelihood ratio (LLR) measure, and the log-area-ratio (LAR) measure which evaluate the spectral envelope of the given disordered speech. Figure 1 shows the objective disordered speech quality assessment block diagram.

3. DYNAMIC TIME WARPING

Conventional objective speech quality measures are used to evaluate the speech quality after speech is coded and decoded or transmitted with noise and channel degradation. In these scenarios, the original high-quality speech and the degraded speech have exactly the same length, which leads to a simple one-to-one comparison of windows from each speech utterance. However, in this project, we use the speech produced

by healthy people as the gold standard to compare with disordered speech. In this case, aligning the two different speech segments to the same reasonable comparable length is crucial. Dynamic time warping (DTW) is the most straightforward solution and is used to solve exactly this problem in speech recognition applications.

Given two speech patterns, $\mathbf{X}$ and $\mathbf{Y}$, these patterns can be represented by a sequence $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{T_x})$ and $(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{T_y})$, where $\mathbf{x}_i$ and $\mathbf{y}_i$ are the feature vectors. As we have noted, in general the sequence of $\mathbf{x}_i$'s will not have the same length as the sequence of $\mathbf{y}_i$'s. In order to determine the distance between $\mathbf{X}$ and $\mathbf{Y}$, given that some distance function $d(\mathbf{x}, \mathbf{y})$ exists, we need a meaningful way to determine how to properly align the vectors for the comparison. DTW is one way that such an alignment can be made [8]. We define two warping functions, ϕ_x and ϕ_y, which transform the indices of the vector sequences to a normalized time axis, k. Thus we have

$$\begin{aligned} i_x &= \phi_x(k), \quad k = 1, 2, \ldots, T, \\ i_y &= \phi_y(k), \quad k = 1, 2, \ldots, T. \end{aligned} \tag{1}$$

This gives us a mapping from $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{T_x})$ to $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T)$ and from $(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{T_y})$ to $(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T)$. With such a mapping, we are able to compute $d_\phi(\mathbf{x}, \mathbf{y})$ using these warping functions, giving us the total distance between two patterns as

$$d_\phi(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^{T} \frac{d(\phi_x(k), \phi_y(k))\, m(k)}{M_\phi}, \tag{2}$$

where $m(k)$ is a path weight and M_ϕ is a normalization factor. Thus, all that remains is the specification of the path ϕ indicated in the above equation. The most common technique is to specify that ϕ is the minimum of all possible paths, subject to certain constraints by using the equation as follows:

$$d(\mathbf{X}, \mathbf{Y}) \simeq \min_{\phi} d_\phi(\mathbf{x}, \mathbf{y}). \tag{3}$$

For time normalization, the optimal path based on DTW has fixed beginning and ending points. Some other constraints may also apply. For example, the path should be monotonic, which requires a positive slope. This constraint eliminates the possibility of reverse warping. Therefore, we choose to enforce the Type III local constraint [8]. In addition, our numerous experimental results show that the eight local constraints will not significantly change the final results. Because of the local continuity constraints, certain portions are excluded from the region the optimal warping path can traverse. By using the maximum and minimum possible path expansion, we can define global path constraints as follows:

$$\begin{aligned} 1 + \frac{(\phi_x(k) - 1)}{Q_{\max}} &\le 1 + Q_{\max}(\phi_x(k) - 1), \\ T_y + Q_{\max}(\phi_x(k) - T_x) &\le T_y + \frac{(\phi_x(k) - T_x)}{Q_{\max}}. \end{aligned} \tag{4}$$

In this aspect, slope weighting along the path adds yet another dimension of control in the search for the optimal warping path. There are four types of slope weighting. The type chosen in this paper is

$$m(k) = \phi_x(k) - \phi_x(k-1) + \phi_y(k) - \phi_y(k-1). \tag{5}$$

If we take the notation $d(i_x, i_y)$ as the distance between $\mathbf{x}_{ix}$ and $\mathbf{y}_{iy}$, which are the elements of $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{T_x})$ and $(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{T_y})$, respectively, and $D(i_x, i_y)$ as the accumulative optimal value, then we can apply the exact local constraint as well as the slope weight to get

$$D(i_x, i_y) = \min \begin{Bmatrix} D(i_x - 2, y_x - 1) + 3d(i_x, i_y) \\ D(i_x - 1, y_x - 1) + 2d(i_x, i_y) \\ D(i_x - 1, y_x - 2) + 3d(i_x, i_y) \end{Bmatrix}. \tag{6}$$

4. OBJECTIVE QUALITY MEASURES

From an anthropomorphic perspective, speech production is very complex but a simple view is that vowels are produced by the lungs, the larynx excitation, and the resonance of the vocal tract. The laryngeal configuration and the tongue's position dramatically change an individual speaker's speech intonation, pitch, or quality. For example, due to differences in tongue positions during pronunciation, nonnative speakers of English may use tongue movements characteristic to their native language, thereby producing a noticeable accent. Similarly, the rigid tongue movement of the Parkinson's patient causes their pronunciation to become distorted. We attempt to develop objective speech quality measures using knowledge of human speech production. However, we first need to define a few terms commonly used in speech processing. A formant is defined as a peak in the speech power spectrum. The pitch of speech is usually determined by the frequency of the excitation signal, which is produced by the vibration of the vocal folds. The vocal tract resonance is usually represented by the spectral envelope.

Some contemporary research has already made progress on objective analyses of disordered speech. For instance, the Computerized Speech Lab (CSL) produced by Kay Elemetrics Corporation is a commercially available hardware and software package for the analysis of disordered speech. The CSL allows a clinician to calculate several measures related to the intelligibility and quality of disordered speech. Another commercial product is the EVA system, made by SQ-Lab, Marseille, France. This system allows simultaneous measurement of acoustic and aerodynamic parameters related to speech production. Acoustic signals are recorded using the microphone built into the pneumotachograph which is used to measure oral airflow. Intraoral pressure may be calculated using a built-in pressure sensor [9]. The majority of such analysis packages allow the calculation of acoustic and aerodynamic parameters such as jitter, shimmer, signal-to-noise ratio, oral airflow, and voice onset time. However, the concordance between these objective measures and perceptual ratings of quality and intelligibility remains at a relatively low

percentage [10, 11], and is often unsuitable for clinical purposes. Many of these measures can only be calculated from relatively steady portions of the speech signal. However, numerous studies have stressed that the unsteady parts of the signal, such as onset, could provide valuable information for objective evaluation of speech and allow finer discrimination of the severity of dysphonia. In addition, many of these measures are calculated from a single vowel that patients are required to produce for a relatively long period of time [12, 13]. In reality, the natural continuous sentence may provide a more accurate picture of the patients speech disorder.

To overcome some of these shortcomings of the existing speech analysis techniques, we propose a new algorithm originally inspired by the speech coding-decoding and speech telecommunications techniques. The first meaningful measure which can be obtained to compare speech differences is to compute the differences of the logarithms of the power spectrum at each frequency range [4]. We use the following equation to represent the difference:

$$d(w) = \ln \left| X(w) \right|^2 - \ln \left| Y(w) \right|^2, \tag{7}$$

where $X(w)$ and $Y(w)$ are the magnitudes in the frequency domain of two compared speech signals. It is also possible to formally express the most easy and straightforward method to stand for the spectral distortion as follows:

$$d(X, Y) = \left(\int_{-\pi}^{\pi} \left| d(w) \right|^k \frac{dw}{2\pi} \right)^{1/k}, \tag{8}$$

where, again, X and Y here represent the two speech signals to be compared.

Although the above method is easy to implement, good results are not guaranteed. Many different types of modified standard objective quality measures have been proposed. These include measures such as the Itakura-Saito (IS) distortion measure, the log-likelihood ratio (LLR) measure, the log-area-ratio (LAR) measure, the segmental SNR measure, and the weighted spectral slope (WSS) measure. In this paper, we chose to investigate the first three measures: IS, LLR and LAR [14, 15, 16].

The IS distortion measure is calculated based on the following equation:

$$d_{\mathrm{IS}}(\mathbf{a}_d, \mathbf{a}_\phi) = \left(\frac{\sigma_\phi^2}{\sigma_d^2} \right) \left(\frac{\mathbf{a}_d \mathbf{R}_\phi \mathbf{a}_d^T}{\mathbf{a}_\phi \mathbf{R}_\phi \mathbf{a}_\phi^T} \right) + \log \left(\frac{\sigma_\phi^2}{\sigma_d^2} \right) - 1, \tag{9}$$

where σ_ϕ^2 and σ_d^2 represent the all-pole gains for the standard healthy people's speech and the test patients' speech. $\mathbf{a}_\phi$ and $\mathbf{a}_d$ are the healthy-speech and patient-speech LPC coefficient vectors, respectively. $\mathbf{R}_\phi$ is the autocorrelation matrix for $x_\phi(n)$, where $x_\phi(n)$ is the sampled speech of healthy people. The elements of $\mathbf{R}_\phi$ are defined as

$$r_\phi(|i-j|) = \sum_{n=1}^{N-|i-j|} r_\phi(n) r_\phi(n + |i-j|), \quad |i-j| = 0, 1, \ldots, p, \tag{10}$$

Table 1: MOS subjective measure evaluation table.

Rating	Speech quality	Level of distortion
5	Excellent	Imperceptible
4	Good	Perceptible, but not annoying
3	Fair	Perceptible, and slightly annoying
2	Poor	Annoying, but not objectionable
1	Unsatisfied	Very annoying and objectionable

where N is the length of the speech frame and p is the order of LPC coefficients.

LLR is similar to the IS measure. However, while the IS measure incorporates the gain factor by using variance terms, LLR only considers the difference between the general spectral shapes. The following equation provides the details for computing the LLR:

$$d_{\mathrm{LLR}}(\mathbf{a}_d, \mathbf{a}_\phi) = \log \left(\frac{\mathbf{a}_d \mathbf{R}_\phi \mathbf{a}_d^T}{\mathbf{a}_\phi \mathbf{R}_\phi \mathbf{a}_\phi^T} \right). \tag{11}$$

LAR is another speech quality assessment measure based on the dissimilarity of LPC coefficients between healthy speech and the patient's speech. Different from LLR, LAR uses the reflection coefficients to calculate the difference and is expressed by the equation

$$d_{\mathrm{LAR}} = \left| \frac{1}{p} \sum_{i=1}^{p} \left(\log \frac{1 + r_\phi(i)}{1 - r_\phi(i)} - \log \frac{1 + r_d(i)}{1 - r_d(i)} \right)^2 \right|^{1/2}, \tag{12}$$

where p is the order of the LPC coefficients, $r_\phi(i)$ and $r_d(i)$ are the ith reflection coefficients of healthy and patient's speech signals.

In the following section describing the experiment and results, we will compare the performances of each of these measures applied to our database. The correlation between these objective quality assessment measures and one subjective quality assessment will also be discussed.

5. SUBJECTIVE QUALITY MEASURES

No matter how speech quality is defined, it must be based on human response and perception. So designing a suitable subjective measure of quality is very important in the assessment of speech quality. Correspondingly, the most important criterion to evaluate the accuracy of an objective measure of quality is to determine its correlation with subjective quality measures.

As discussed in Section 1, subjective measures can be broadly divided into utilitarian and analytic categories. Without loss of generalization, we will use two of the utilitarian methods for our investigation. One reliable and easily implemented subjective utilitarian measure is the mean opinion score (MOS) [1, 4]. In this method, human listeners rate the speech under test on the five-point scale shown in Table 1. Related research shows that as few as five but no more than nine categories are enough for the assessment of

Table 2: Moderate-severe subjective measure evaluation table.

Rating	Level of distortion
3	Moderate
2	Moderate to severe
1	Severe

quality. The final speech quality assessment value can be calculated as the average of the responses of several listeners. The MOS test is widely used in the telecommunications area to compare the original signal quality with that of the distorted signal. For disordered speech analysis, however, it may not be feasible to categorize sentences as "perceptible, but not annoying" or "annoying, but not objectionable." Therefore, a different commonly used subjective utilitarian measure was obtained. In this test, listeners rated the sentences into three categories: mild, moderate, or severe [5, 6, 7]. A similar 4-point rating scale, called the GRBAS method, has been presented for the evaluation of disorder voice quality [17]. In these subjective tests, each test sentence was assigned a score based on whether the disordered sentence quality was perceived to be mild, moderate, or severe. Based on our database of Parkinson's patients tested in this experiment, we modified the mild-moderate-severe rating scale to have three new levels: moderate, moderate to severe, and severe. The details and criteria for these ratings are listed in Table 2. The following procedures were followed when obtaining perceptual judgment in the present experiment: Listeners were asked to listen carefully to each test sentence. Listeners were allowed to hear the test sentence as many times as needed to ensure that they assigned the most appropriate score to each sentence. Listeners were asked to read the criteria table (Tables 1 and 2) carefully and were required to assign a score to each sentence based on the level of distortion described in the tables.

6. EXPERIMENTAL RESULTS

The speech database used in this experiment was collected by the experimenters at the Motor Movement Disorders Clinic, University of Florida. Ten patients with Parkinson's disease were recorded reading a standard passage ("Grandfather Passage"). Additionally, the same passage was also recorded from four healthy adult speakers. Although speakers vary in their rate of speech, this passage takes approximately 1 minute to read. Three successive sentences (around 15 seconds in duration) were selected from this passage for acoustic and perceptual analyses. The sentences include "You wish to know all about my grandfather. Well, he is nearly ninety three years old. He dresses himself in an ancient black frock coat, usually minus several buttons." The fourteen speakers were divided into two groups—males and females. In the first listening test, six listeners evaluated the speech of four Parkinson's patients and one healthy speaker. In the second listening test, we tested twelve listeners who rated the speech of seven Parkinson's patients and one healthy speaker. Of the 18 participants in the listening tests, six were from the USA, five from China, five from India, one from Korea, and one from Turkey. Seven of them were male and the rest were female. All listeners spoke fluent English.

The first listening test was used to obtain ratings using the MOS criteria listed in Table 1. Listeners gave an individual score to each sentence. In this study, two different methods were used to compare the objective and the subjective measures. In the first method, all MOS scores given by the listeners were correlated with the distance measures calculated by the various algorithms. In the second approach, the order of the MOS scores (rather than the actual value of the MOS scores) was correlated with the distance measures. In this approach, listeners simply ordered each sentence from the best to the worst quality. If two or more sentences were given the same rank, listeners were asked to listen carefully and choose different ranks for each sentence. In contrast, in the first method, listeners may end up giving identical integer scores to two speech segments even though one may sound noticeably better than the other. Table 3 gives the details on all the sentences scored using MOS scale for male speakers only. Sentences labelled as P1, P2, P3, and P4 were spoken by the Parkinson's patients and H1 is the sentences spoken by the healthy speaker. The six listeners are labelled as List1 to List6.

One sentence from a healthy speaker was used as the standard sentence for calculating the objective measures of quality. DTW was first applied to align this standard sentence with each patient's sentence. Figure 2 shows the optimal frame match path between the standard healthy speech and the patient's speech. For the second method, every procedure is the same except for replacing the exact score by the relative order. Therefore, in Table 3, each column is the order given by each listener.

Finally, the three distortion measures (IS, LLR, and LAR) were calculated. The last three columns in Table 3 show the exact values of IS, LLR, and LAR, respectively. In Table 4, the last three columns show the relative order of the distortion scores obtained from each speaker. Figure 3 shows the healthy speech waveform (upper panel), the patient speech waveform (middle panel) and their distortion curve calculated by the IS measure (lower panel). Figure 4 shows a similar comparison based on LLR and Figure 5 shows the same comparison based on LAR. Figure 6 exhibits the histogram of the distortion values, which may give us deeper insight about the differences between the healthy speaker and the patient's speech. This may provide greater information than the use of a single number obtained by averaging the distortion measures across a number of frames.

As discussed earlier, the quality of an objective measure is determined by how well it predicts the subjective measure. The following formula is widely used to evaluate the performance of objective measures:

$$\hat{\rho} = \frac{\sum_d (S_d - \bar{S}_d)(O_d - \bar{O}_d)}{\left(\sum_d (S_d - \bar{S}_d)^2 \sum_d (O_d - \bar{O}_d)^2\right)^{1/2}}, \tag{13}$$

where S_d and O_d are subjective and objective results. $\bar{S}_d$ and $\bar{O}_d$ are their corresponding average values. Table 3 shows all

Table 3: Subjective test results and their correlation with objective test using method 1 in the first round.

Subject	List1	List2	List3	List4	List5	List6	Avg.	IS	LLR	LAR
P1	2	3	2	2	3	2	2.33	71 035	197.5	1441.5
P2	2	1	2	1	2	1	1.50	769 990	175.6	1054.2
P3	3	1	1	1	2	2	1.67	572 200	152.3	1014.9
P4	3	2	3	2	3	3	2.67	304 150	218.8	1025.4
H1	5	5	5	5	5	5	5	24 155	96.2	752.5
Corr.	—	—	—	—	—	—	—	0.7638	0.6419	0.5729

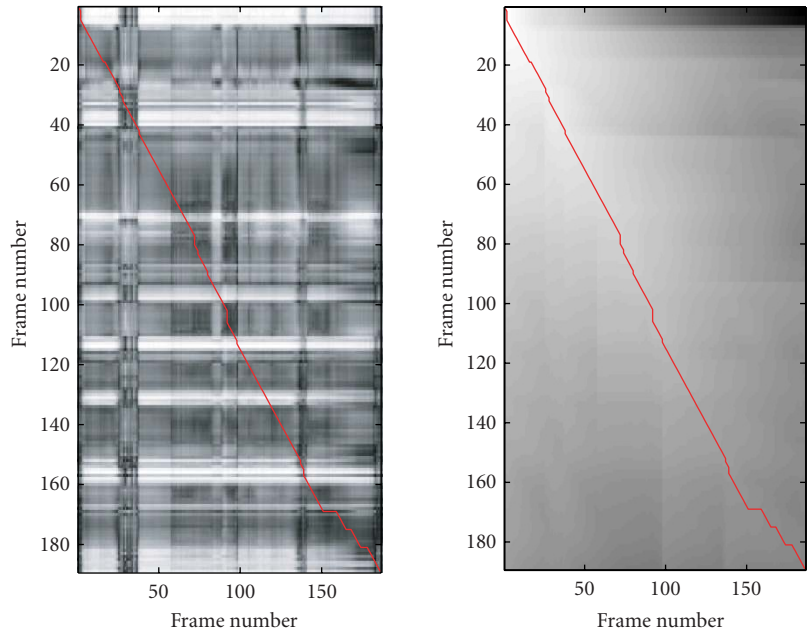

Figure 2: Dynamic time warping (DTW) optimal path between the recorded speech of a healthy person (horizontal axis) and a Parkinson's patient (vertical axis).

Table 4: Subjective test results and their correlation with objective test using method 2 in the first round.

Subject	List1	List2	List3	List4	List5	List6	Avg.	IS	LLR	LAR
P1	5	2	3	3	3	3	3.2	2	4	5
P2	4	4	4	5	4	5	4.3	5	3	4
P3	2	5	5	4	5	4	4.2	4	2	2
P4	3	3	2	2	2	2	2.3	3	5	3
H1	1	1	1	1	1	1	1	1	1	1
Corr.	—	—	—	—	—	—	—	0.8684	0.1828	0.5142

three objective measures and their correlation values based on method 1. The IS measure, with a correlation of 0.7638, showed the best performance. Table 4 lists the correlation values based on method 2, and once again the IS measure showed the highest correlation of 0.8684. In analyzing (9), (11) and (12), we can see that the good performance of the IS measure might be partially due to the fact that it not only considers the general spectral difference, but also uses the variance term to take into account the gain factor of the all-pole filter model.

After completing the preliminary test, a second test was conducted to validate our conclusion that IS is a good measure of disordered speech quality. In this test, speech samples from a larger number of patients with Parkinson's disease (seven instead of four) were rated by more listeners (twelve instead of six). In addition to the MOS scores, listeners were also asked to categorize the speech samples as Normal, moderate, moderate to severe, or severe. To highlight the validity of the IS measures, only this measure was calculated for the speech samples used in the second test. Table 5 shows the

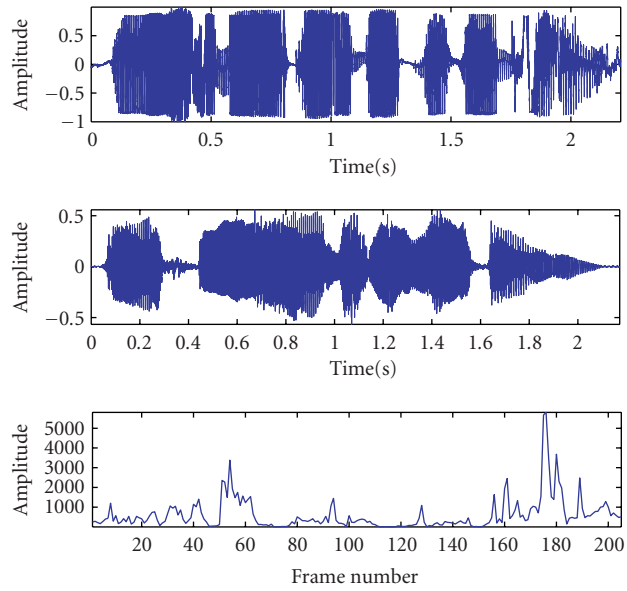

Figure 3: IS value (lower) versus healthy speech waveform (upper) and patient speech waveform (middle).

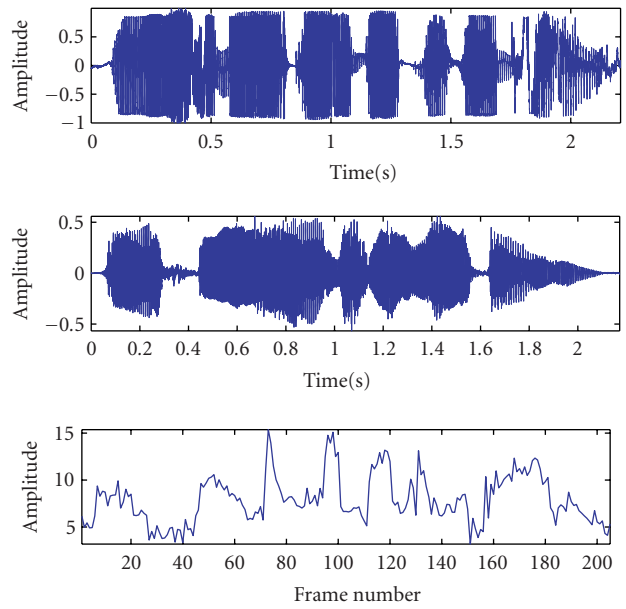

Figure 4: LLR value (lower) versus healthy speech waveform (upper) and patient speech waveform (middle).

MOS from individual listeners, the average MOS, and the correlation between the IS measure and MOS values based on method 1 described earlier. This correlation was found to be 0.8032 and is comparable with 0.7638 obtained in the first round test. Table 6 shows the moderate-severe test scores from each listener, the average moderate-severe test scores, and the correlation between the IS measure and the subjective ratings. Once again, a correlation of 0.7417 was obtained which is comparable to that obtained in the first round test.

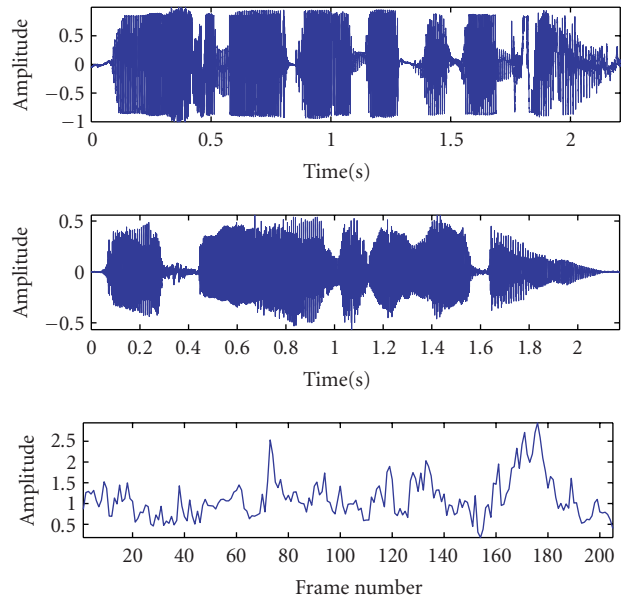

Figure 5: LAR value (lower) versus healthy speech waveform (upper) and patient speech waveform (middle).

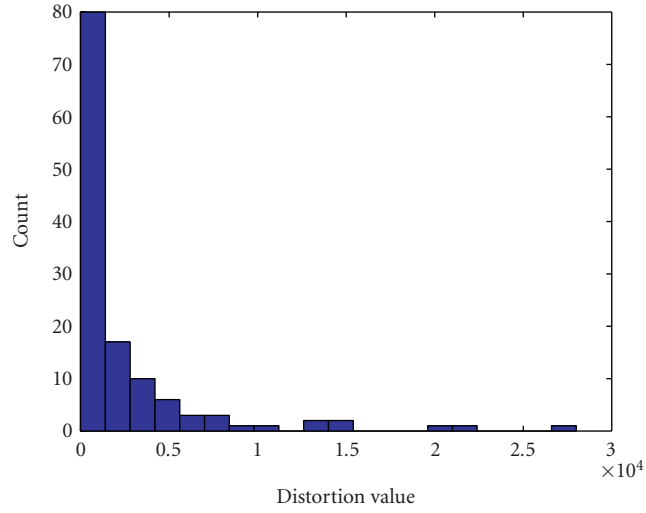

Figure 6: The histogram of the distortion values based on the IS method.

All objective speech quality assessment criteria (IS, LLR, LAR etc.) proposed above mainly focus on the speech spectral envelope. From the perceptual point of view, we are mainly interested in how to efficiently evaluate the speech intelligibility and quality. However, intelligibility and quality are not the only aspects of the overall speech quality evaluation. Many other factors that affect speech quality may also need to be considered. For instance, Hansen and Nandkumar proposed that pitch turbulence (PT) may be used to evaluate the monotone or pitch variation, which is directly related to the laryngeal excitation signal. Similarly, energy turbulence (ET) is another important factor used to evaluate

Table 5: Subjective test results and their correlation with objective test using method 1 in the second round based on MOS test.

Subject	List1	List2	List3	List4	List5	List6	List7	List8	List9	List10	List11	List12	Avg.	IS
P1	3	2	4	3	2	2	3	3	3	3	1	1	2.50	41 500
P2	2	3	3	2	3	1	2	2	2	3	2	1	2.17	84 200
P3	1	2	2	1	2	1	2	1	1	2	1	1	1.42	264 000
P4	4	4	4	4	4	3	5	4	5	4	4	4	4.08	10 300
P5	4	4	3	3	3	4	5	4	4	5	4	4	3.92	29 800
P6	1	3	2	1	2	2	3	1	2	3	2	1	1.92	205 000
P7	2	3	2	2	2	3	4	3	3	3	3	2	2.67	103 000
H1	5	5	5	5	5	3	5	5	5	5	5	5	4.83	6010
Corr.	—	—	—	—	—	—	—	—	—	—	—	—	—	0.8032

Table 6: Subjective test results and their correlation with objective test using method 1 in the second round based on moderate-severe test.

Subject	List1	List2	List3	List4	List5	List6	List7	List8	List9	List10	List11	List12	Avg.	IS
P1	1	2	2	1	2	1	2	1	1	1	2	1	1.42	205 000
P2	2	2	1	2	2	2	2	2	3	2	2	2	2	103 000
P3	3	3	3	3	3	3	3	3	3	3	3	3	3	10 300
P4	1	1	1	1	1	1	1	2	1	1	1	1	1.08	264 000
P5	2	2	2	1	2	2	2	1	2	2	1	1	1.67	84 200
P6	2	3	1	2	2	2	3	2	2	2	2	2	2.08	41 500
P7	3	3	3	3	3	3	3	3	3	3	3	3	3	29 800
Corr.	—	—	—	—	—	—	—	—	—	—	—	—	—	0.7417

the monoloudness or energy variation [2]. The following equations give the exact mathematic expressions for these measures:

$$\mathrm{PT} = \frac{1}{N-1}\sum_{i=1}^{N-1} \left| P(i+1) - P(i) \right|,$$
$$\mathrm{ET} = \frac{1}{N-1}\sum_{i=1}^{N-1} \left| E(i+1) - E(i) \right|, \tag{14}$$

where N is the total number of frames of the given sentence, and $P(i)$ and $E(i)$ represent the pitch and energy of the frame i. We used the data obtained in the first round of evaluation to test the correlation between these measures (PT and ET) and the subjective ratings. Table 7 shows the pitch turbulence (PT) and energy turbulence (ET) values calculated from (14) as well as their correlation based on method 1. Table 8 shows the similar results based on the method 2. Figures 7 and 8 show the pitch turbulence and energy turbulence from a given speech signal. Based on Tables 7 and 8, it appears that PT and ET are poorly correlated with the subjective assessments, using either method 1 or method 2. This suggests that during subjective assessment, humans put most of their emphasis on intelligibility, which, from a signal processing view, is related primarily to the spectral envelope. The excitation (pitch) and energy variation are not as important as spectral envelope variation in the perception of overall speech quality. Even in our current algorithm, pitch and energy turbulence were not very efficient in predicting the overall speech quality. Potentially, even though the correlation performance with one-dimensional evaluation (such as MOS) is poor, these two parameters may correlate well with multidimensional evaluation (such as DAM).

Table 7: Subjective test results and their correlation with PT and ET test using method 1.

Subject	Avg.	PT	ET
P1	2.33	13.2777	3.9040
P2	1.50	8.0712	8.0712
P3	1.67	4.5775	11.9782
P4	2.67	16.3607	5.8966
H1	5	4.2815	8.3446
Corr.	—	0.1264	0.1137

Table 8: Subjective test results and their correlation with PT and ET test using method 2.

Subject	Avg.	PT	ET
P1	3.2	4	5
P2	4.3	3	3
P3	4.2	2	1
P4	2.3	5	4
H1	1	1	2
Corr.	—	0.1828	0.0800

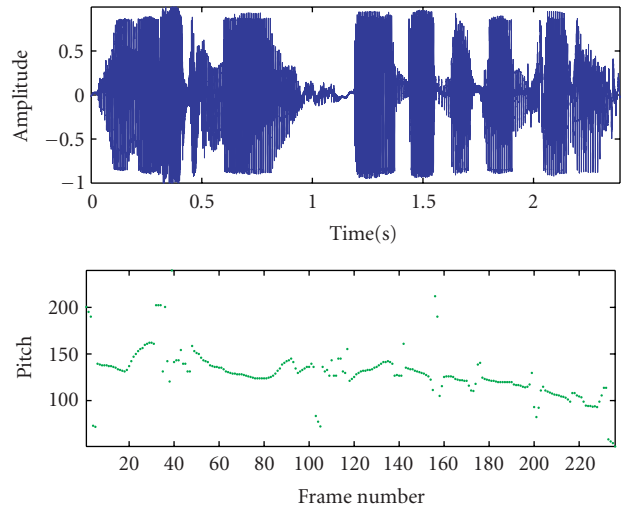

Figure 7: The pitch turbulence (lower) from a given speech signal (upper).

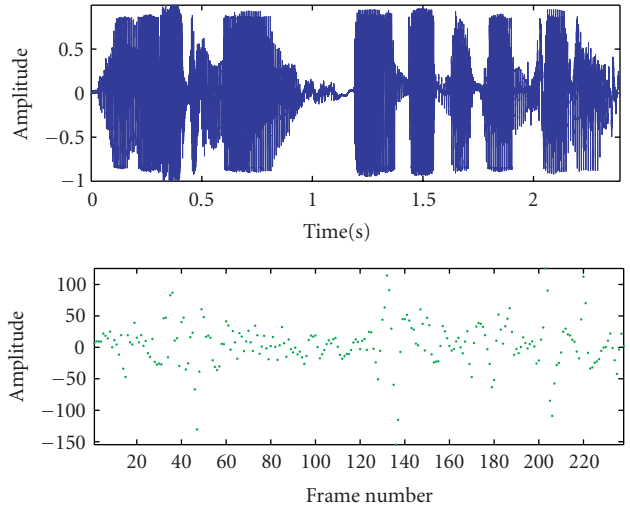

Figure 8: The energy turbulence (lower) from a given speech signal (upper).

7. CONCLUSION

Objective evaluation of disordered speech quality is not an easy task. In this paper, we discuss three objective quality assessment measures and one subjective measure. By evaluating our speech database, the IS measure showed a strong correlation with the MOS tests. Therefore, the IS measure is suggested to be more suitable than LLR and LAR for use as a reliable tool to evaluate the overall quality of disordered speech. The IS measure could also be used to predict the subjective quality measure MOS score given by humans.

ACKNOWLEDGMENT

The authors are grateful to three reviewers who provided us with a large number of detailed suggestions for improving the submitted manuscript.

REFERENCES

[1] S. Quanckenbush, T. Barnwell, and M. Clements, *Objective Measures of Speech Quality*, Prentice Hall, New York, NY, USA, 1988.

[2] J. Hansen and S. Nandkumar, "Objective quality assessment and the RPE-LTP vocoder in different noise and language conditions," *Journal of the Acoustical Society of America*, vol. 97, no. 1, pp. 609–627, 1995.

[3] J. Hansen and L. Arslan, "Robust feature-estimation and objective quality assessment for noisy speech recognition using the credit card corpus," *IEEE Trans. Speech Audio Processing*, vol. 3, no. 3, pp. 169–184, 1995.

[4] S. Dimolitsas, "Objective speech distortion measures and their relevance to speech quality assessments," *IEE Proceedings Part I: Communications, Speech and Vision*, vol. 136, no. 5, pp. 317–324, 1989.

[5] L. Ramig, C. Bonitati, J. Lemke, and Y. Horii, "Voice treatment for patients with Parkinson disease: Development of an approach and preliminary efficacy data," *Journal of Medical Speech-Language Pathology*, vol. 2, no. 3, pp. 191–209, 1994.

[6] S. Countryman, L. Ramig, and A. Pawlas, "Speech and voice deficits in Parkinsonian Plus syndromes: Can they be treated?" *Journal of Medical Speech-Language Pathology*, vol. 2, no. 3, pp. 211–225, 1994.

[7] S. Countryman and L. Ramig, "Effects of intensive voice therapy on voice deficits associated with bilateral thalamotomy in Parkinson disease: A case study," *Journal of Medical Speech-Language Pathology*, vol. 1, no. 4, pp. 233–250, 1993.

[8] L. Rabiner and B. Juang, *Fundamental of Speech Recognition*, Prentice Hall, New York, NY, USA, 1984.

[9] P. Yu, M. Ouaknine, J. Revis, and A. Giovanni, "Objective voice analysis for dysphonic patients: a multiparametric protocol including acoustic and aerodynamic measurements," *Journal of Voice*, vol. 15, no. 4, pp. 529–542, 2001.

[10] A. Giovanni, D. Robert, N. Estublier, B. Teston, M. Zanaret, and M. Cannoni, "Objective evaluation of dysphonia: preliminary results of a device allowing simultaneous acoustic and aerodynamic measurements," *Folia Phoniatr Logop*, vol. 48, no. 4, pp. 175–185, 1996.

[11] J. Revis, A. Giovanni, F. Wuyts, and J. Triglia, "Comparison of different voice samples for perceptual analysis," *Folia Phoniatr Logop*, vol. 51, no. 3, pp. 108–116, 1999.

[12] D. Berry, K. Verdolini, D. Montequin, M. Hess, R. Chan, and I. Titze, "A quantitative output-cost ratio in voice production," *Journal of Speech, Language and Hearing Research*, vol. 44, no. 1, pp. 29–37, 2001.

[13] P. Dejonckere, C. Obbens, G. De Moor, and G. Wieneke, "Perceptual evaluation of dysphonia: Reliability and relevance," *Folia Phoniatr Logop*, vol. 45, no. 2, pp. 76–83, 1993.

[14] E. Wallen and J. Hansen, "A screening test for speech pathology assessment using objective quality measures," in *Proc. 4th International Conference on Spoken Language Proceedings (ICSLP '96)*, vol. 2, pp. 776–779, Philadelphia, Pa, USA, October 1996.

[15] L. Thorpe and W. Yang, "Performance of current perceptual objective speech quality measures," in *Proc. IEEE Workshop on Speech Coding Proceedings (SCW '99)*, pp. 144–146, Porvoo, Finland, June 1999.

[16] J. Hansen and B. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," On-line Technical Report.

[17] M. Hirano, *Psycho-Acoustic Evaluation of Voice: GRBAS Scale for Evaluating the Hoarse Voice*, Springer Verlag, New York, NY, USA, 1981.

Lingyun Gu received his B.S. and M.S. degrees in electrical engineering from the University of Electronic Science and Technology of China (UESTC) and Old Dominion University in 1998 and 2002, respectively. He is currently pursuing his Ph.D. in the Computational NeuroEngineering Laboratory, Electrical and Computer Engineering Department, University of Florida (UF). His main research interests are in robust speech recognition, speech signal processing, and auditory production and perception.

John G. Harris received his B.S. and M.S. degrees in electrical engineering from MIT in 1983 and 1986. He earned his Ph.D. degree from Caltech in the interdisciplinary Computation and Neural Systems Program in 1991. After a two-year postdoc at the MIT AI Lab, Dr. Harris joined the Electrical and Computer Engineering Department, University of Florida (UF). He is currently an Associate Professor and leads the Hybrid Signal Processing Group in researching biologically inspired circuits, architectures, and algorithms for signal processing. Dr. Harris has published over 100 research papers and patents in this area. He codirects the Computational NeuroEngineering Laboratory and has a joint appointment in the Biomedical Engineering Department at UF.

Rahul Shrivastav earned his B.S. degree in 1995 and M.S. degree in 1997 in speech and hearing sciences from the University of Mysore, India. He completed his Ph.D. degree in speech and hearing science from Indiana University, Bloomington, in 2001. Currently he is on the faculty at the Department of Communication Sciences and Disorders, University of Florida. His research is studying the factors that affect the perception of voice quality and speech intelligibility in patients with a variety of speech disorders.

Christine Sapienza received her Ph.D. degree in speech science from The State University of New York at Buffalo in 1993. Currently, she is a Professor in the Department of Communication Sciences and Disorders, University of Florida. Her most recent work has focused on the use of strength training paradigms in multiple populations including voice disorders, Parkinson's disease, spinal cord injury, and multiple sclerosis. She maintains an active research laboratory with 7 current Ph.D. students. Her clinical work takes place at Ayers Outpatient Voice Clinic and the Motor Movement Disorders Clinic at the University of Florida. She also is a Research Health Scientist at the Brain Rehabilitation Research Center, Malcom Randall VA, Gainesville, Florida.