# DISORDERED SPEECH EVALUATION USING OBJECTIVE QUALITY MEASURES

†*Lingyun Gu* †*John G. Harris* ‡*Rahul Shrivastav* ‡*Christine Sapienza*

†Department of Electrical and Computer Engineering
‡Department of Communication Sciences and Disorders
University of Florida, Gainesville, FL, U.S.A.

**ABSTRACT**

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.

## 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 the 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 a few 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. 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 often 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 subjective measures. 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 characterized by an impairment in one or more of the systems involved in speech. 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. 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 re-

duced 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 parameters that can represent 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.

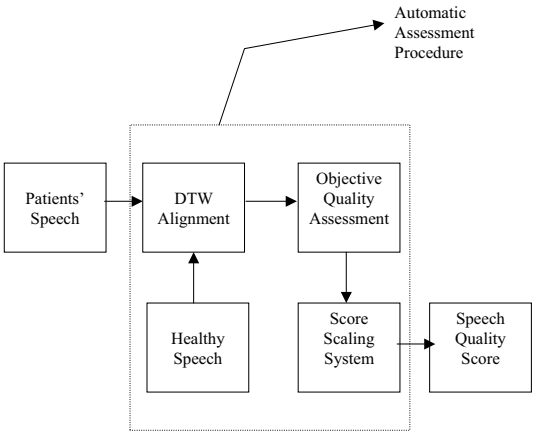


**Fig. 1**. Objective patients' speech quality assessment block diagram.

| 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 |

**Table 1**. MOS Subjective Measure Evaluation Table.

| Rating | Level of Distortion |
|---|---|
| 3 | Moderate |
| 2 | Moderate to Severe |
| 1 | Severe |

**Table 2**. Moderate-Severe Subjective Measure Evaluation Table.

## 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. Full details about the DTW technique can be found in [5].

## 4. OBJECTIVE QUALITY MEASURES

Some contemporary research has already made progress on objective analyses of disordered speech. For instance, the Computerized Speech Lab (CSL) produced by Kay Elemetrics Corp. and the EVA system made by SQ-Lab, Marseille, France. 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 relatively low, and are often unsuitable for clinical purposes. To overcome some of these shortcomings of existing speech analysis techniques, we propose a new algorithm originally inspired by speech coding/decoding and speech telecommunications techniques. Here we deploy three modified measures to compute the objective score of disordered speech after applying DTW. These include the following measures: Itakura-Saito (IS) Distortion Measure, Log-Likelihood Ratio (LLR) Measure and Log-Area-Ratio (LAR) Measure [6], [7], [8].

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

$$d_{IS}(a_d, a_\phi) = (\frac{\sigma_\phi^2}{\sigma_d^2})(\frac{a_d R_\phi a_d^T}{a_\phi R_\phi a_\phi^T}) + \log(\frac{\sigma_\phi^2}{\sigma_d^2}) - 1 \qquad (1)$$

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

$$r_\phi(|i-j|) = \sum_{n=1}^{N-|i-j|} r_\phi(n) r_\phi(n+|i-j|), |i-j| = 0, 1, \ldots, p \qquad (2)$$

where $N$ is the length of the speech frame and $p$ is the order of the 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:

$$d_{LLR}(a_d, a_\phi) = \log(\frac{a_d R_\phi a_d^T}{a_\phi R_\phi a_\phi^T}) \qquad (3)$$

LAR is another speech quality assessment measure based on the

| | List1 | List2 | List3 | List4 | List5 | List6 | Avg | IS | LLR | LAR |
|---|---|---|---|---|---|---|---|---|---|---|
| P1 | 2 | 3 | 2 | 2 | 3 | 2 | 2.33 | 71035 | 197.5 | 1441.5 |
| P2 | 2 | 1 | 2 | 1 | 2 | 1 | 1.50 | 769990 | 175.6 | 1054.2 |
| P3 | 3 | 1 | 1 | 1 | 2 | 2 | 1.67 | 572200 | 152.3 | 1014.9 |
| P4 | 3 | 2 | 3 | 2 | 3 | 3 | 2.67 | 304150 | 218.8 | 1025.4 |
| H1 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 24155 | 96.2 | 752.5 |
| Corr | | | | | | | | .7638 | .6419 | .5729 |

**Table 3**. Subjective test results and their correlation with objective test in the first round

| | 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 | 41500 |
| P2 | 2 | 3 | 3 | 2 | 3 | 1 | 2 | 2 | 2 | 3 | 2 | 1 | 2.17 | 84200 |
| P3 | 1 | 2 | 2 | 1 | 2 | 1 | 2 | 1 | 1 | 2 | 1 | 1 | 1.42 | 264000 |
| P4 | 4 | 4 | 4 | 4 | 4 | 3 | 5 | 4 | 5 | 4 | 4 | 4 | 4.08 | 10300 |
| P5 | 4 | 4 | 3 | 3 | 3 | 4 | 5 | 4 | 4 | 5 | 4 | 4 | 3.92 | 29800 |
| P6 | 1 | 3 | 2 | 1 | 2 | 2 | 3 | 1 | 2 | 3 | 2 | 1 | 1.92 | 205000 |
| P7 | 2 | 3 | 2 | 2 | 2 | 3 | 4 | 3 | 3 | 3 | 3 | 2 | 2.67 | 103000 |
| H1 | 5 | 5 | 5 | 5 | 5 | 3 | 5 | 5 | 5 | 5 | 5 | 5 | 4.83 | 6010 |
| Corr | | | | | | | | | | | | | | .8032 |

**Table 4**. Subjective test results and their correlation with objective test in the second round based on MOS test.

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:

$$d_{LAR} = |\frac{1}{p}\sum_{i=1}^{p}(\log\frac{1+r_\phi(i)}{1-r_\phi(i)} - \log\frac{1+r_d(i)}{1-r_d(i)})^2|^{1/2} \qquad (4)$$

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

## 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 criteria to evaluate the accuracy of an objective measure of quality is to determine its correlation with subjective quality measures.

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 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 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 subjective utilitarian measure is proposed in this paper. 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 (Table 1 and Table 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 at the University of Florida. Ten patients with Parkinson's disease were recorded reading a standard passage (the "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 total 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

| | 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 | 205000 |
| P2 | 2 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 3 | 2 | 2 | 2 | 2 | 103000 |
| P3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 10300 |
| P4 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1.08 | 264000 |
| P5 | 2 | 2 | 2 | 1 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 1.67 | 84200 |
| P6 | 2 | 3 | 1 | 2 | 2 | 2 | 3 | 2 | 2 | 2 | 2 | 2 | 2.08 | 41500 |
| P7 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 29800 |
| Corr | | | | | | | | | | | | | | .7417 |

**Table 5**. Subjective test results and their correlation with the objective test in the second round based on Moderate-Severe test.

criteria listed in Table 1. Listeners gave an individual score to each sentence. All MOS scores given by the listeners were correlated with the distance measures calculated by the various algorithms. Sentences labelled as P1, P2, P3 and P4 were spoken by the Parkinson's patients and H1 is the label for the sentences spoken by the healthy speaker. The six listeners are labelled as Lis1 to Lis6. 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. 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.

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:

$$\hat{\rho} = \frac{\sum_d (S_d - \bar{S}_d)(O_d - \bar{O}_d)}{(\sum_d (S_d - \bar{S}_d)^2 \sum_d (O_d - \bar{O}_d)^2)^{1/2}} \quad (5)$$

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 three objective measures and their correlation values. The IS measure, with a correlation of 0.7638, showed the best performance. In analyzing equations 1, 3 and 4, we can see that the good performance of the IS measures might be partially due to the fact that it not only considers the general spectral difference, but 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 the IS measure 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-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 4 shows the MOS from individual listeners, the average MOS and correlation between the IS measure and MOS values. This correlation was found to be 0.8032 and is comparable with 0.7638 obtained in the first round test. Table 5 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.

## 7. CONCLUSION

Objective evaluation of disordered speech quality is not an easy task. In this paper, we discuss three objective quality assessment measures and two subjective measures. By evaluating our speech database, the IS measure showed a strong correlation with the subjective 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 measures' score given by humans.

## 8. REFERENCES


[1] S. Quanckenbush, T. Barnwell and M. Clements, *Objective Measures of Speech Quality*. Prentice Hall, U.S.A, 1988.

[2] J. Hansen and S. Nandkumar, *Objective Quality Assessment and the RPE-LTP Vocoder in Different Noise and Language Conditions*. The Journal of Acoustic Society of America, Vol.97, No.1, pp.609-627, January 1995.

[3] J. Hansen and L. Arslan, *Robust Feature-Estimation and Objective Quality Assessment for Noisy Speech Recognition Using the Credit Card Corpus*. IEEE Transactions on Speech and Audio Processing, Vol.3, No.3, pp.169-184, May 1995.

[4] S. Dimolitsas, *Objective Speech Distortion Measures and Their Relevance to Speech Quality Assessments*. IEE Proceedings, Vol.136, No.5, pp.317-324, October 1989.

[5] L. Rabiner and B. Juang, *Fundamental of speech recognition*. Prentice Hall, U.S.A, 1984.

[6] E. Wallen and J. Hansen, *A Screening Test for Speech Pathology Assessment Using Objective Quality Measures*. ICSLP Proceedings, Vol.2, pp.776-779, October 1996.

[7] L. Thorpe and W. Yang, *Performance of Current Perceptual Objective Speech Quality Measures*. IEEE Workshop on Speech Coding Proceedings, pp.144-146, June 1999.

[8] J. Hansen and B. Pellom, *An Effective Quality Evaluation Protocol for Speech Enhancement Algorithms*. On-line Technical Report.