A COMPARISON OF TECHNIQUES FOR THE MEASUREMENT OF INTELLIGIBILITY IN HYPOKINETIC DYSARTHRIA

By

TRACI L. REYNOLDS

A THESIS PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MATER OF ARTS
UNIVERSITY OF FLORIDA

2012

To all those diagnosed with Parkinson’s disease

ACKNOWLEDGMENTS

I would like to thank my family, friends, and colleagues who have made these past years memorable for me. Mark, Supraja, Nour, Tony, Isaac, Lisa, and Karessa were all helpful with various aspects of my data collection and analysis and were great company at lunch-time! My parents have shown unending support in every endeavor I have pursued and they never cease encouraging me to follow my passions, wherever they may lead me. Donna and Georges have been constant companions. I greatly appreciate all your help day to day; your care and love mean the world to me.

I would like to thank my committee members, Dr. John Rosenbek and Dr. Judith Wingate, for their contributions to various aspects of this work. Thanks go out to Dr. Rosenbek who inspires me to strive for excellence in the clinic and beyond, and to Dr. Wingate whose remarkable love and compassion toward each student is a treasure to be truly cherished; thank you for all of your support! I would especially like to thank Dr. Rahul Shrivastav for his mentorship during the course of my experience at the University of Florida throughout my undergraduate and graduate studies. He is a great researcher, a smart leader, and a wonderful person. I am grateful to have had the opportunity to work under his guidance.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ASR	Automatic speech recognition
DAF	Delayed auditory feedback
DBS	Deep-brain stimulation
DME	Direct Magnitude Estimation
EBP	Evidence-based practice
EMST	Expiratory muscle strength training
HD	Hypokinetc dysarthria
IPD	Idiopathic Parkinson’s disease
LSVT	Lee-Silverman voice therapy
MEP	Maximum expiratory pressure
MIP	Maximum inspiratory pressure
PD	Parkinson’s disease
VHI	Voice Handicap Index
VOT	Voice onset time
VRQL	Voice-relate quality of life

Abstract of Thesis Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Master of Arts

A COMPARISON OF TECHNIQUES FOR THE MEASUREMENT OF INTELLIGIBILITY IN HYPOKINETIC DYSARTHRIA

By

Traci L. Reynolds

August 2012

Chair: Judith Wingate
Major: Communication Sciences and Disorders

The development of an automated measurement tool which could predict the average judgments of naïve listeners would be optimal for patient diagnostics and treatment in order to optimize specificity and reliability of intelligibility measurement.

The current study seeks to evaluate the impact of intelligibility estimation tasks and stimulus familiarity upon intelligibility scores in hypokinetic dysarthria. Specifically, question 1 inquires about the impact of familiarity upon DME intelligibility data. Question 2 will investigate the correlation between two dysarthria measurement paradigms, DME and Transcription. Additionally, this work will help create baseline perceptual data that may be used for the development of automatic measures of speech intelligibility in patients with dysarthria subsequent to PD.

Two primary experiments have been conducted. The first experiment included the training and extensive testing of 10 listeners on a DME task as a measure of speech intelligibility of talkers diagnosed with PD. The second experiment consisted of 100 listeners participating in a one-time, ten minute transcription task in which they were presented with just 15 different sentences from 15 different talkers.

Results:

Question 1: A sentence effect is evident particularly in talkers with lower intelligibility. The more familiar constant sentence was consistently rated as more highly intelligible than the random sentences.

Question 2: There is a moderate positive correlation between the DME and Transcription data. There is higher variability between the tasks in talkers with moderate dysarthria.

CHAPTER 1
INTRODUCTION

Parkinson's Disease

Parkinson's disease (PD) is a neurodegenerative disease which initially impacts one's movement and may eventually degrade the integrity of cognition, sensation, and behavior (Rusz, Cmejla, Ruzickova, & Ruzicka 2010). The cause is yet unclear, though possible genetic and environmental triggers have been identified. However advanced age remains the prime predisposing factor for PD. This disorder occurs when the neurons of the substantia nigra, which trigger dopamine release, are damaged or destroyed. This in turn reduces the amount of dopamine released, beyond that typical in normal aging, depriving the system (Hoehn & Yahr, 1967). Lack of dopamine has numerous implications for the motor system including rigidity, tremor, akinesia, bradykinesia, postural instability and imbalance, as well as reduced range of motion and difficulty initiating movements (Duffy, 2005). These motor difficulties may manifest within the bulbar system and translate into possible impairment of voice, speech, and swallow function.

Speech and Voice Characteristics of PD

Impaired motor-speech functioning associated with Parkinson's disease is called hypokinetic dysarthria (HD). The defining characteristics of this dysarthria include low respiratory drive and loudness, articulatory imprecision and reduced range of motion in the articulators, rapid rate of speech, reduced range of pitch and inflection, and inappropriate pausing (Darley, Aronson, & Brown, 1969a). The neuro-motor deficits associated with this disease have logically been paralleled with subsequent bulbar impairments, for example bradykinesia and akinesia have been implicated in the

increased time required to initiate lip and tongue movement (Goberman & Coelho, 2002). However, Ali et al. (1996) found no relation between limb tremor and lingual tremor, nor between muscular rigidity and dysmotility of the pharyngeal wall. Additionally, severity and degree of motor involvement in PD do not necessarily correlate with severity of voice or speech dysfunction (Sarno, 1968). Metter and Hanson (1986) attempted to quantify five characteristics that were rated on a scale of 1-7, resulting in dysarthria scores ranging from 5 to 35. They found that bradykinesia, rigidity, and facial motility had no relationship to total dysarthria scores and that likewise, no relationship was observed between dyskinesia, tremor, or duration of Parkinsonism and perceived severity of dysarthria. However a drawback of this dysarthria measure is that it was not weighted by the impact of each speech characteristic upon intelligibility. For example, nasality and imprecise articulation were each rated 1 to 7, though articulation would more greatly impact intelligibility. An understanding of speech anatomy and physiology, as well as how intelligibility may be impacted by various breakdowns, is necessary in the clinical assessment of dysarthria.

Hypokinetic Dysarthria Assessment

Comprehensive clinical voice assessment for voice and speech disorders associated with Parkinson's disease is multi-faceted, including surveys and acoustic measures, as well as respiratory and articulatory tasks. The voice of a patient with Parkinson's disease may include qualities such as hoarseness, roughness, tremulousness, softness and breathiness. These talkers may also present with reduced loudness, hypernasality, and reduced range of pitch (Duffy, 2005). The collection of acoustic measures including noise to harmonic ratio, fundamental frequency and frequency range, and dynamic range of loudness is more common clinically and may be

helpful in diagnosis. These data can also be used as a baseline against which progression of the disease or improvements with therapy may be measured. Additionally, voice onset time (VOT) and VOT ratio as well as jitter and shimmer are other notable acoustic measures in the literature which may not be as accessible in a clinical setting (Rusz, et al. 2010). Tracking loudness may be advantageous, as a "consistently greater level of progressive intensity decay compared with matched controls" (p. 109) was observed in patients with PD (Ho, Iansek, & Bradshaw, 2001). Oguz, et al. found that differences in jitter, loudness, and harmonic to noise ratios were significantly different in people with PD compared to age- and gender-matched controls; therefore acoustic profiles may also be deemed useful in diagnosis (Oguz, et al., 2006). Fischer and Goberman (2010) support the usefulness of examining both VOT and VOT ratio in individuals with PD in order to dissociate between rate-related VOT changes and true VOT changes (Fischer & Goberman, 2010).

Additionally, respiratory and articulatory measures are important to assessment and treatment. Maximum expiratory and possibly inspiratory pressure, maximum phonation time, and diadochokinetic rate should also be analyzed in assessment to gauge respiratory drive and oropharyngeal aspects of speech (Duffy, 2005).

Hypokinetic Dysarthria Treatment

Treatment of voice disorders associated with Parkinson's disease is not widely sought nor implemented, with only 3-4% of the nearly 90% of PD patients with voice and speech problems ever seeking rehabilitative assistance (Trail, et al., 2005), possibly due to the degenerative nature of the disease or lack of evidence based practice (EBP) supported therapeutic techniques. Lee-Silverman Voice Treatment (LSVT) is quickly becoming a standard technique of choice as it is supported by numerous group studies

with clinically and statistically significant data demonstrating the efficacy of its impact upon impaired respiratory and speech mechanisms. The intensive nature of this therapy is highly taxing financially, physically, mentally, temporally; but it improves respiratory function and even generalizes beyond increasing loudness to “improve swallowing, articulation, communicative gestures, facial expression, and neural functioning” (Sapir, Ramig, & Fox, 2008, p. 207). Participants are required to speak as though shouting in order to compensate for the reduced loudness within the population as well as to aid in the recalibration of speakers who may have lost sensitivity in appropriately gauging verbal output. However, Frost, et al. (2010) propose that patients are able to perceive their difficulty accurately due to their finding of high correlation between Voice Handicap Index (VHI) scores from patients with PD and their intelligibility scores as rated by listeners both before and after deep brain stimulation (DBS) surgery (Frost, Tripoliti, Hariz, Pring, & Limousin, 2010).

A second popular therapeutic paradigm is Expiratory Muscle Strength Training (EMST) to enhance respiratory drive by increasing the load upon the muscles of expiration with the use of a pressure release valve. Participants breathe a strong blast of air into a device with a valve which will only open if the set pressure has been met, forcing the release of the valve. Generally, the required expiratory pressure is set at 75% of the participant’s maximal expiratory pressure in order to tax the respiratory muscles without “maxing out” on every trial. This pressure is adjusted as needed, according to the participant’s progress. While this is also often a 4 week program, as with LSVT, this training is less challenging to manage logistically. Increases are seen in maximum phonation times, maximum expiratory pressure, and loudness, resulting in the

consequent reduction of breathy or harsh vocal quality (Sapienza, Troche, Pitts, & Davenport, 2011).

Additional programs intended to target loudness and possibly prosody include singing, masking and delayed auditory feedback, as well as transcranial magnetic stimulation. A singing protocol researched by Di Benedetto, et al. found significant improvement of residual capacity, maximal expiratory pressure (MEP), maximal inspiratory pressure (MIP), maximum duration of vowel phonation, and even enhanced prosody while reading following participation in their entertaining therapy sessions (Di Benedetto, et al., 2009). Masking was studied by Quedas, et al. who found that a Lombard effect occurred in both PD and control participants in the same way (Quedas, Duprat Ade, & Gasparini, 2009). The Lombard effect, or reflex, refers to the tendency of talkers to alter acoustic features of their output in relation to environmental contexts, for example, increasing the loudness of one's output in order to adapt to a noisy situation. Masking effects were compared with delayed auditory feedback (DAF) and amplification, revealing "improvement in vocal quality, increase in loudness and overall strain level" (p. 219) in the masking situation; a decrease in loudness and overall strain was observed in the DAF and amplification situations, as well as reduced speech rate and articulatory precision (Coutinho, Diaféria, Oliveira, & Behlau, 2009). Therefore, masking was found to be more highly efficacious than DAF or amplification in the treatment of HD.

Influence of PD Treatment upon Voice

Medical treatment of Parkinson's Disease is still developing, though a few strategies have proven valuable in restoration of motor function; however the impact of these developments in the realm of voice and speech remain unclear (Trail, et al.,

2005). The use of deep brain stimulation (DBS) has become more refined over time and is a popular treatment for the motor impacts of Parkinson's disease. The use of DBS resulted in significant improvement of gross motor disabilities, but failed to have a significant influence upon voice and speech function (Valálik, Smehák, Bognár, & Csókay, 2011). One study did find that harmonic to noise ratio improved in repetition during DBS (Van Lancker Sidtis, Rogers, Godier, Tagliati, Sidtis, 2010).

L-Dopa may be administered as a replacement for depleted dopamine and has positive effects on motor function, though no significant impact upon voice has been found in "on vs. off" conditions (Plowman-Prine, et al., 2009). Another medical treatment is compensation for the non-closure of the glottis (Midi, et al., 2008). Injection of Cymetra was analyzed by Sewall, et al. who found that, "collagen injection in patients with (Parkinson's-related dysphonia) is safe, well tolerated, and is an effective temporary method of subjectively improving voice and speech in selected patients with IPD" (Sewall, et al., 2006, p.1740). It is important to take into account the impact of medical treatments and interventions upon the variation of voice and speech production in this population.

Intelligibility

Intelligibility is the degree to which a speaker's message is accurately recovered by the listener (Kent et al., 1989). This concept is solely influenced by impaired speech production, impaired transmission of the utterance, or some listener variable which inhibits his or her perception of the stimulus. In HD, the break down in intelligibility resides within the speech production phase. Reduced intelligibility impacts communicative efficacy, and in turn often affects the quality of life maintained by individuals with intelligibility deficits. In the HD population, intelligibility is frequently

judged by clinicians as an indicator of dysarthric severity and also as a baseline against which to measure declines related to disease progression as well as therapy gains. The use of intelligibility scores as a functional communication measure was supported by Beukelman and Yorkston (1979) when they compared the relationship between intelligibility scores and information transfer. Numerous techniques have been developed for the quantification of speech intelligibility.

Quantifying Intelligibility

Obtaining precise speech intelligibility judgments has important implications for both research and clinical purposes. The ability to accurately quantify the degree of impairment may allow for the description of speech characteristics associated with certain diseases or syndromes, enable clinicians to estimate the communicative handicap of patients, and facilitate the measurement of treatment effects as well as degenerative advances of disease progression (Beukelman & Yorkston, 1979). Current methods of quantifying intelligibility both clinically and in research include percentage estimates, transcription tasks, and scaling techniques including rating scales and direct magnitude estimation. Of clinicians who implement some method of quantifying speech intelligibility, 75% utilize percent estimation of intelligible words and 55% used a rating scale system as part of standard assessment (Beukelman & Yorkston, 1980). Each paradigm offers distinct advantages and disadvantages. As such the selection of which measure to utilize is influenced by the type of information the researcher or clinician requires.

In order to develop a functional model of dysarthric speech perception, it is necessary to primarily ensure that listener judgments are both sensitive and reliable. Intelligibility judgments of a single talker may vary greatly across listeners and even

within listeners across presentations. This is a notable concern due to the reality that judgments in a clinical setting are frequently made by a single clinician (Beukelman & Yorkston, 1980). In order to enhance reliability, “at a minimum, the speech samples must be unfamiliar to the judges, and a number of judges are necessary to reduce the inter-judge reliability problems” (Beukelman & Yorkston, 1980, p.41). However, even with numerous naïve listeners, responses across trials are often highly inconsistent; therefore, averaging across numerous naïve listeners and trials will improve rating predictability (Shrivastav, Sapienza, & Nandur, 2005). Of course, such a method is time consuming and logistically does not lend itself to a clinical setting. Each intelligibility quantification method currently available is characterized by both positive and negative attributes. The development of a new technique which diminishes these negative aspects while maintaining or enhancing the positive would be ideal. Popular techniques include percentage estimates, transcription, linear scaling techniques, and magnitude estimation.

Percentage Estimates

This technique is a subjective measure by which listeners estimate the percent of verbal output which they understand from a talker. According to Buekelman and Yorkston’s survey, percentage estimates of speech intelligibility are the most popular rating system implemented among clinicians. Variability both within and across listeners is commonly a fundamental challenge due to the nature of subjective judgments. Multiple factors influence the degree of variation with this task including, but not limited to: experience level, listener bias, passage familiarity, familiarity with the talker, concept of range of extremes along the continuum being measured, as well as contextual cues and recency effects (Beukelman & Yorkston, 1980). Beukelman and Yorkston (1980)

found that passage and talker familiarity did not impact intelligibility judgments for mild and severe dysarthria samples, but in patients with moderate dysarthria, intelligibility percentage estimates varied to a significant degree as a result of familiarity. Therefore the use of such estimates “should be employed with reservation” (p.41), particularly when making judgments about talkers with moderate intelligibility deficits.

Transcription Tasks

Phonetic transcription involves listeners converting acoustic speech signals into written graphemes. This transcription is then compared with the target phonemes in the utterance in order to calculate errors and estimate intelligibility (Tikofsky & Tikofsky, 1964; Yorkston & Beukelman, 1980). This technique is primarily utilized in research as a presumably objective measure; however, as with other intensive listening tasks, it may also be susceptible to variability resulting from effects of fatigue, attention, memory, experience or training. Yorkston and Beukelman (1980) found that transcription tasks produce less variability than rating scales.

Transcription and information transfer are highly correlated (Beukelman & Yorkston, 1979), however this technique is time consuming and requires multiple naïve listeners in order to produce unbiased intelligibility estimates. For this reason, this method is often considered unpractical for daily clinical use.

Linear Scaling Techniques

This is another subjective technique for which listeners assign a number along a linear continuum as representative of the degree to which the feature being judged is present. For example, 1-5, 1-7, and 1-10 scales are common. A majority of clinicians and experiments implement unanchored rating scales to quantify listener perceptions. Such a model is inherently dependent upon the experience or memory of the listeners.

As with percentage estimates, variability both within and across listeners is commonly a challenge due to a number of variables including training, familiarity and recency effects as well as the influence of stimuli number, type, and scale range. Additionally, listener judgments may be impacted by momentary alterations in attention, memory, level of fatigue, and other chance factors (Poulton, 1989). Some of this variability may be reduced by randomizing presentation of stimuli across trials. Kreiman, Gerratt, Kempster, Erman, and Berke (1993) found that the variability in rating scale estimates was greatest for stimuli with an average rating in the middle of the scale and less at the two extremes, signifying that the use of a scaling paradigm may not be the most appropriate option for quantifying speech intelligibility of moderately impaired talkers, including many patients with PD.

Additionally, these tasks are built upon the assumption that listeners will be able to rate speech in reference to constant perceptual distances from other presented stimuli. Stevens (1946) argued that the intervals on the linear rating scale do not actually represent equal intervals, but might be curvilinear in nature. Such a relation rather signifies that speech as a percept be understood and analyzed as a prothetic continuum.

Magnitude Estimation

Magnitude estimation requires listeners to assign a number relative to the magnitude of the target feature perceived. Due to the prothetic nature of the human perception of many sensations, including brightness, pain, and speech quality, it follows that a rating scale developed to quantify speech intelligibility need not be represented on a linear, but rather a ratio scale. Stevens made popular the use of ratio judgments along a nearly unlimited scale in the realms of numerous psychophysical measures,

resulting in a logarithmic relationship between the physical and perceptual magnitude of the stimulus known as Steven's Law; however, few studies of speech perception have implemented such a scale.

Like in the above techniques, direct magnitude estimation (DME) is susceptible to the same listener biases, including centering bias, familiarity and recency effects as well as the influence of stimuli number, type, and scale range. The resultant inter-listener variation common to this technique makes it impractical to compare results from different studies. Many studies also utilize different standard or modulus stimuli to appropriately calibrate listener judgments. The use of a standard may aid in reducing the impact of some of the listener biases mentioned above; however, this may also reduce the ability to generalize the data set in order to compare with other studies.

Summary and Purpose

Currently utilized methods of intelligibility measurement which are clinically feasible are lacking in specificity or reliability, and those methods which are most highly specific and reliable are not easily carried out in contexts outside of research due to the large number of naïve listeners required. Therefore the development of an automated measurement tool which could predict the average judgments of naïve listeners would be optimal for patient diagnostics and treatment.

Ultimately, automated measures of intelligibility will be developed using automatic speech recognition technology and signal distortion measures. The thousands of judgments from trained and naïve listeners collected in this research will be utilized in training algorithms to best match average listener responses. Once developed, these algorithms will serve as an automated measurement tool which can predict the average listener judgments of speech intelligibility with novel stimuli.

A large database of speech samples from hundreds of people diagnosed with PD will be created. Preliminary analysis of the resulting database will be conducted to assess variability in intelligibility across sentence stimuli. This will be done to ensure that the database consists of a wide range of dysarthric severity as this database could be used in the development of automatic speech recognition systems and as an ideal baseline for test development.

Two experiments will be completed. The first experiment will consist of two steps, including training of listeners to utilize DME as a measure of speech intelligibility, followed by more extensive experimentation with a fraction of trained listeners. The goal of this experiment is to establish baseline perceptual data for the development and/or training of any automatic system of speech intelligibility measurement. The second experiment will consist of one-hundred listeners participating in a one-time, ten minute transcription task in which they are presented with just 15 different sentences from 15 different talkers. This data will help identify the differences between DME and transcription scores of intelligibility and provide two different kinds of baselines for developing any automated measures of speech intelligibility.

Overall, the current study seeks to evaluate the impact of intelligibility estimation tasks and stimulus familiarity upon intelligibility scores in dysarthria. Specifically, question 1 inquires about the impact of familiarity upon DME intelligibility data. Question 2 will investigate the correlation between two dysarthric measurement paradigms, DME and Transcription. Additionally, this work will help create baseline perceptual data that may be used for the development of automatic measures of speech intelligibility in patients with dysarthria subsequent to PD.

CHAPTER 2
METHODS

There were three phases to this study. First, a database of speech recordings from people diagnosed with Parkinson's disease was created. Next, a small group of listeners were trained to rate the intelligibility of these samples in a direct magnitude estimation task. Following training, a portion of the trained listeners participated in an additional direct magnitude estimation task, scaling the perceived intelligibility of the Parkinson's speech samples. Finally, one-hundred listeners, naïve to dysarthric speech, were recruited for participation in a short transcription task.

Development of PD Speech Database

Participants

Participants diagnosed with Parkinson's disease were recruited from numerous Parkinson's support groups throughout Gainesville, FL and surrounding communities. A total of 180 volunteer participants were recruited ranging in age from 44 to 92, with a mean age of 71 years. Male participants totaled 122, and 58 females participated. Information collected from each talker included: age, gender, veteran status, and diagnosis of Parkinson's disease.

Procedure

Prior to testing, volunteers were briefed on the nature of the experiment as well as the expectations for participants. Once any questions or concerns were addressed, both the participant and researcher signed the informed consent form. Each talker produced five vowel sounds (/a/, /i/, /o/, /u/, /ae/), 50 words, 15 sentences, and finally The Rainbow Passage (Appendix A). Participants were instructed to read each word or sentence as an independent utterance to avoid list reading.

A different randomized list of words was developed for each talker from the 500 word database used by Yorkston and Beukelman (1980). One word was randomly selected from each of the 50 lists of ten words in the database, to create unique 50-word lists. The 15 sentences were selected from the Speech Perception In Noise (SPIN) lists developed by Kalikow, Stevens, and Elliott (1977). Highly predictable sentences were used. These stimuli were selected due to the balanced phonetic nature of the sentences. All word and sentence stimuli were presented individually to the talkers via a laptop screen. Each PowerPoint slide displayed one word or sentence to be spoken by the participant, then the researcher advanced to the next slide until all stimuli were spoken The vowels were presented orally and participants were asked to sustain each phonation for a minimum duration of five seconds. These were produced at a comfortable loudness and pitch. A segment of The Rainbow Passage (Fairbanks, 1960) was presented on a large-type printout which talkers read aloud at a comfortable rate and loudness.

Table 2-1. List of recorded vocal tasks elicited from participants with Parkinson's disease.

Task Code	Speech Data
[Task 1]	Vowels /a/, /i/, /o/, /u/, /ae/ sustained approximately 5 seconds at a comfortable pitch and loudness.
[Task 2]	50 words randomly selected for each participant from a pool of 500 words. Presented individually on laptop screen and read aloud.
[Task 3]	Read 15 sentences aloud. Presented one at a time on laptop screen.
[Task 4]	Read 'The Rainbow Passage', standard phonetically balanced text.

Equipment

All spoken stimuli were recorded in one sitting using a head-mounted microphone (Audiotechnica, ATM21a) that was connected to a Marantz solid state digital recorder (PMD671) which digitized the samples at 44.1 kHz. Each speech sample was screened to ensure that no recording errors, including peak clipping and inappropriate environmental noise, were present. In the case of a recording error, participants were immediately instructed to repeat the given stimulus. Samples with other recording errors, such as peak clipping which cannot necessarily be caught on site, were not used in listening tasks. Finally, the speech sample from each talker was segmented into 71 wav files (5 vowels, 15 sentences, 50 words, one passage).

Direct Magnitude Estimation

The direct magnitude estimation (DME) paradigm was utilized in order to best represent and understand the psychoacoustic nature of speech intelligibility in talkers with hypokinetic dysarthria. Listeners were trained to rate the intelligibility of a portion of the samples developed in the database in a direct magnitude estimation task. Later, a portion of the trained listeners participated in an additional direct magnitude estimation task, scaling the perceived intelligibility of the Parkinson's speech samples.

Listener Training

Participants

Twenty-eight naïve listeners majoring in Speech, Language and Hearing Sciences or Linguistics underwent training to familiarize them with the perceptual scaling procedure and to develop their concept of intelligibility. The mean age of the listeners was 20.5 years, ranging from 18 years to 24 years. One of the participants was male and 27 were female. Listeners were all native speakers of American English with no

previous history of speech problems. Each participant passed a hearing screening bilaterally using air-conduction pure-tone audiometer at 20dB HL for 250Hz, 500Hz, 1000Hz, 2000Hz, and 4000Hz. Listeners were given $5 per hour as compensation for participation in the experiment.

Equipment

The training session was conducted in a group context at a quiet computer lab reserved for this use. Stimuli were presented through headphones at a comfortable listening level and participant responses were input via keyboard and computer.

Stimuli

A “training set” of 500 stimuli spanning the range of speech impairments in patients with PD was presented. Fifty talkers were selected to be used in the training task. Ten sentences were randomly selected from each of the 50 talkers. Attention was given to ensure that the entire range of intelligibility within the database was well represented by those selected.

Procedure

Prior to testing, listeners were briefed on the nature of the experiment as well as the expectations for participants. Once any questions or concerns were addressed, both the participant and researcher signed the informed consent form. Intelligibility as a concept was explained, as was the nature of the direct magnitude estimation task. Intelligibility judgments were made using direct magnitude estimation. On the scale, each number represents the ratio of intelligibility across samples. For example, a stimulus perceived as having a normal degree of intelligibility may be rated 100, while a stimulus perceived to be twice as unintelligible would receive a rating of 200, and so on.

A “normal” speech anchor was presented every twenty trials with a standard rating of 100 in order to calibrate listeners to the scale. While standards are often a reference for the center of a particular scale, judgments are highly dependent upon the characteristics of that standard (Weismer & Laures, 2002). Therefore, the current study anchored the “normal” end of the scale with a standard referent at the lower extreme.

Listeners were provided short breaks periodically in order to maintain an optimal level of attention and to minimize fatigue. This rating task was completed in one 1-2 hour session.

Reliability

Each listener’s judgments on the training task were tabulated and correlated with the group mean score for each stimulus and any listener whose intelligibility ratings correlated with the group mean at less than $r = 0.70$ were considered to have failed the training session. Listeners who failed the primary training session had the opportunity to complete a second training session; however if their responses between the first and second training revealed a low intra-judge correlation of less than $r = 0.70$, the listener was excluded from the study. Two participants had poor intra-judge reliability and were not included in further testing.

DME Extended Listening Task

Participants

Ten trained listeners from the above training task participated in the direct magnitude estimation task. All but two of the trained listeners were invited to participate in further testing, and data from the first ten to respond and participate was used. The mean age of the listeners was 20.8 years, ranging from 18 years to 23 years. All 10 participants in this task were female. Listeners were given $5 per hour as compensation

for participation in the experiment. An additional $10 bonus was paid following the completion of the experiment in order to quell attrition. There was some attrition due to one participant being unable to complete her second testing session.

Equipment

Listeners were tested individually in a double-walled, sound treated booth. SykoFizX software was used to present stimuli in the right ear monaurally at 75 dB SPL using the RP2 or RX6 processor (Tucker-Davis Technologies, Inc.) with ER-2 ear inserts. Responses were input via computer and keyboard.

Stimuli

Eighty talkers from the Parkinson's speech database were selected to be used in the current experiment. Attention was given to ensure that the entire range of intelligibility within the database was well represented by those selected. The direct magnitude estimation task entailed making intelligibility judgments for two sentences from each of the selected 80 talkers. One sentence was the same across all talkers (sentence 13) and the second sentence was randomly selected from the remaining 14 sentences (Appendix A). The use of sentence 13 was selected due to the fact that this sentence was read toward the end of the recording session and it has the highest number of syllables; therefore this sentence may be more representative of the talkers' speech in fatigue. Sentence 13 will be referred to as the constant sentence, or Sentence C, and the remaining 14 sentences will be referred to as Random Sentence, or Sentence R.

Procedure

Prior to testing, listeners were briefed on the nature of the experiment as well as the expectations for participants. Once any questions or concerns were addressed, both

the participant and researcher signed the informed consent form. Listeners were presented with two sentences from each of the 80 talkers and were instructed to choose an intelligibility rating for each stimulus on the same magnitude estimation scale used in the training paradigm (1-1000 ratio scale). A single trial consisted of all 160 stimuli being randomly presented. A total of 5 trials were completed for each listener over two separate testing days. The presentation of all stimuli was randomized across listeners and trials. As in the training protocol, the same “normal” speech anchor was presented every twenty stimuli with a standard rating of 100 in order to maintain calibration of listeners to the scale. Listeners were provided short (3 to 5 minute) breaks periodically in order to maintain an optimal level of attention and to minimize fatigue. This rating task was completed in two 1-2 hour sessions.

Transcription

Participants

One-hundred and four naïve listeners were recruited at a highly trafficked library at the University of Florida over the course of one month. Inclusionary criteria for participation in this task included being a native speaker of American English in addition to having no previous history of speech, language, or hearing problems. Information collected from participants was limited to age and gender. The mean age of the listeners was 22.6 years, ranging from 18 years to 55 years of age. Forty-eight males and 56 females participated in the transcription experiment. Listeners were given $5 as compensation for participation in the experiment.

Equipment

The transcription task was conducted in a quiet study room in a campus library. One to two participants were tested simultaneously. Stimuli were presented binaurally

through Windows Media Player via HD 570 headphones. Listeners typed each sentence they heard into their respective Excel spreadsheet using a keyboard and laptop.

Stimuli

Sixty sentences from the eighty talkers utilized in the above DME testing were randomly selected to be used in the current experiment. Only three talkers were represented twice in the random sample, and some talkers were not incorporated into this task due to the randomized nature of stimuli selection. The 60 stimuli were randomized into four groups of 15 sentences, such that no grouping repeated the same talker or sentence. This was done to ensure that the naïve listeners did not gain familiarity with a talker or with the sentence stimuli. Further, as depicted in Table 2-2, the grouping of the 60 stimuli was randomized four additional times, for a total of five sets of four groups, or 20 different sequences of 15 sentences. This was done to neutralize effects of presentation order, task familiarity, and other unknown effects. Each participant listened to one group, 15 sentences.

Each of the 20 different sequences was developed into wav files including the 15 randomized sentences, which were graded, in addition to two practice sentences. The

Table 2-2. The 20 stimuli wav files were developed by dividing the 60 sentences into 4 groups of 15. The 60 sentences were randomized four additional times, creating 5 sets of 4 groups each.

Wav File #	Set #	Group #	Wav File #	Set #	Group
1	1	1	11	3	3
2	1	2	12	3	4
3	1	3	13	4	1
4	1	4	14	4	2
5	2	1	15	4	3
6	2	2	16	4	4
7	2	3	17	5	1
8	2	4	18	5	2
9	3	1	19	5	3
10	3	2	20	5	4

two practice sentences were presented at the beginning of each sequence as the first and second sentences to be transcribed. The practice sentences were judged by the researchers to be highly intelligible and were distinct from each other, not overlapping with any of the randomized PD stimuli. The two practice sentences were pulled from a different database of patients from an Ear, Nose, and Throat practice in Gainesville, FL which was developed using identical recording procedure; however, more sentence stimuli were recorded for this database. The practice sentences were included in order to adjust listeners to the task as well as to check technical factors, including adequate volume level.

Procedure

Prior to testing, listeners were briefed on the nature of the experiment as well as the expectations for participants. Once any questions or concerns were addressed, both the participant and researcher signed the informed consent form. Each participant listened to one of the 20 wav files described above. Each wav file was presented to five separate listeners and presentation was determined sequentially. For example, the first participant listened to wav file one and the 21st participant also listened to wave file one.

Listeners were asked to type what they understood in the corresponding sentence number on the Excel spreadsheet following each sentence presentation, during the 10 second window of silence. Participants were instructed to type exactly what was heard, making a guess if the stimulus was moderately unintelligible, and even leaving a blank word, phrase, or entire sentence if intelligibility was very poor. Each of the 17 sentences per wav file was presented only once. Following completion of all 17 sentences, the listener was asked to review the spreadsheet for typos in order to ensure that what was saved corresponded exactly with what he or she understood. Participation required a

time-commitment of 10-15 minutes. Data collection took place at a similar time of day, in the afternoon.

Grading. The transcription task was graded by counting one point for each syllable of content words in every sentence as detailed in the grading template (Appendix C); therefore possible points per sentence were not constant across stimuli due to varied word lengths as well as different amounts of content words. Each participant spreadsheet was graded by the researcher in order to avoid problems such as point deduction for mere typographical errors, which may have altered averages had an automated paradigm been implemented. Grading was completed twice in order to ensure appropriate scoring and to minimize effects of human error.

Percentage scores were derived for each sentence/talker combination from the total points earned divided by total points possible, multiplied by 100. A total of 15 percentage scores were recorded from each listener and compared across the four other participants who heard the same wav file- the same 15 sentences, in the same order. The percentage scores were averaged across the five listeners for each of the 15 sentences, resulting in 15 group percentage scores. Finally, the five different group scores for each sentence/talker were averaged, resulting in a mean percentage score for each sentence/talker. Talkers were rank ordered from least to greatest percentage of intelligibility scores across listeners and trials for analysis.

CHAPTER 3
RESULTS

Reliability

Direct Magnitude Estimation

Intra-judge and inter-judge reliability measures were analyzed for each listener, across talkers, and between sentences. Ratings from the standard sentence used as an anchor were not factored into analysis. Reliability for the direct magnitude estimation judgments was calculated with the Log_{10} of the absolute magnitude estimates because these judgments had been made on a ratio scale.

The average Pearson's correlation for intra-judge reliability across the five trials was 0.66 (standard deviation: 0.289, range: 0.01-0.92). Figure 3-1 plots the minimum and maximum intra-judge correlations within the five trials for each listener as well as their mean and median correlations (top and bottom of the boxes, respectively). The five trials from each listener were averaged for each stimulus, reducing the data to one

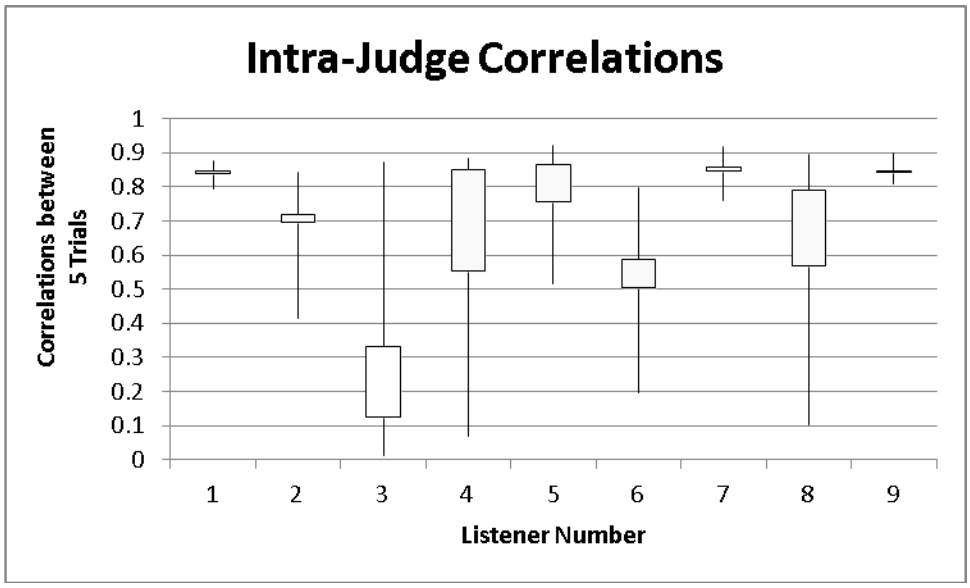

Figure 3-1. The minimum, maximum, mean (upper point of box), and median (bottom of box) intra-judge correlations for each listener.

average response from each listener for each of the 160 stimuli.

Inter-judge reliability for the rating scale was determined by calculating the Pearson's correlation across listeners. The correlation of inter-judge reliability across the nine listeners was significant with an average Pearson's r equal to 0.76, $p<0.01$ (standard deviation: 0.06, range: 0.63-0.86).

The first and last trials were averaged across listeners for each talker in order to analyze familiarity effects. These curves are displayed in Figure 3-2. The average absolute difference between the mean response across listeners for each talker from Trial 1 and Trial 5 was 30.74 with a standard deviation of 28.85 and ranging from a minimum absolute difference of 0.556 and a maximum absolute difference of 198.13. A Pearson's correlation comparing the average ratings of trials 1 and 5 was conducted, with $r=0.93$, $p<0.01$.

Figure 3-3 details how each listener used the scale with minimum and maximum

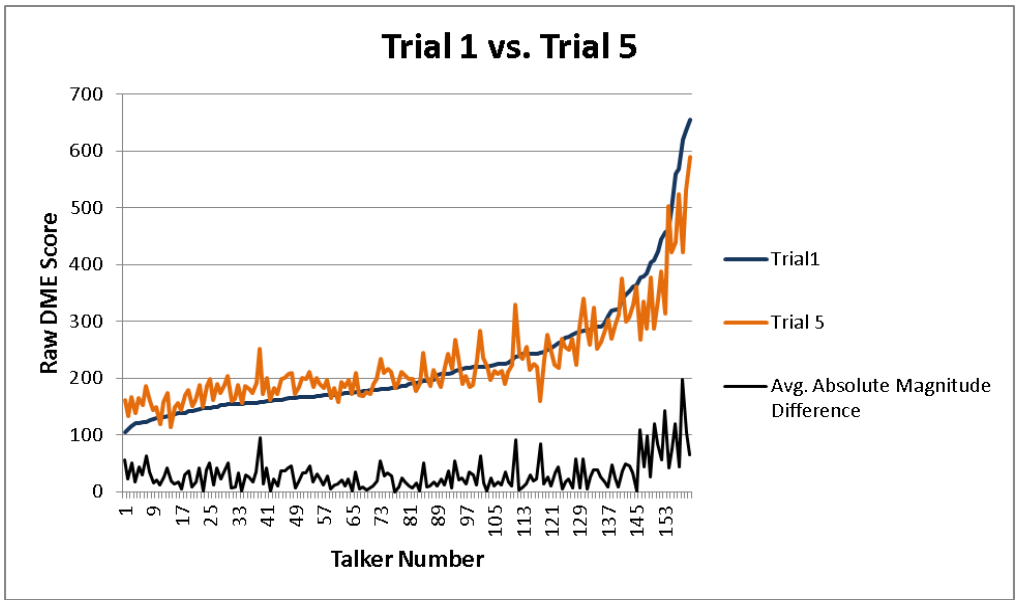

Figure 3-2. Raw DME Intelligibility judgments of Trial 1 and Trial 5, ordered from least to greatest score on Trial 1. ($r=0.93$, $p<0.01$)

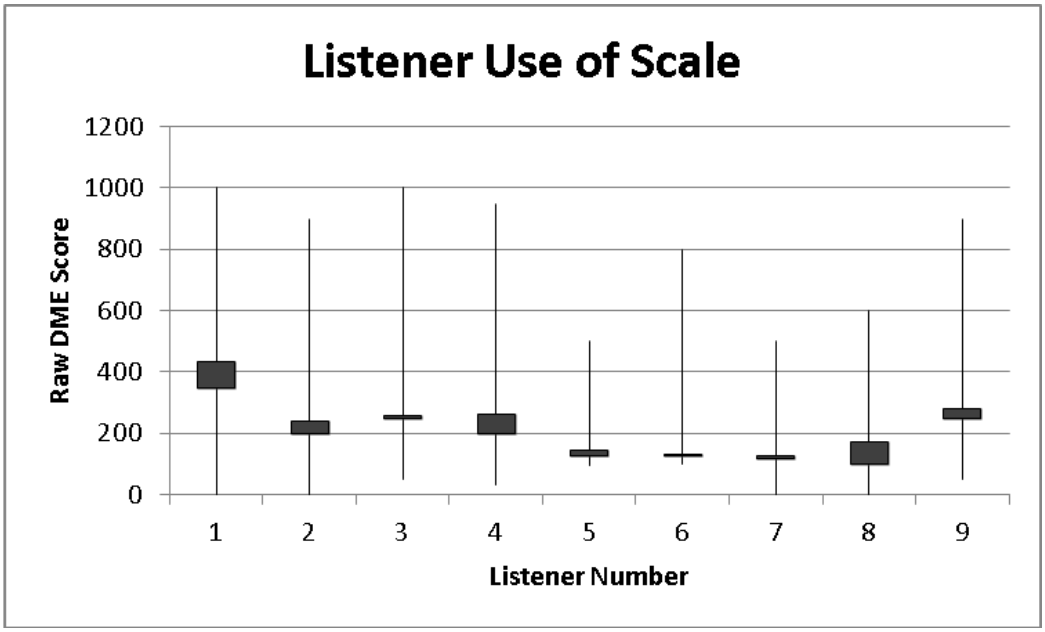

Figure 3-3. The minimum, maximum, mean (upper point of box), and median (bottom of box) DME ratings for each listener.

judgment ratings as well as the mean and median ratings from each listener, represented by the top and bottom of each box, respectively. For example, Listener one expands the entire 1-1000 range with a mean rating of ~436. While other listeners may also have implemented a wider range of the scale, most sustained a lower mean.

Transcription

Intra-judge analysis was completed across the five listeners who were presented with exactly the same stimuli in the same order utilizing Pearson's correlation. These correlations were then averaged across groups ($r=0.61$, $p<0.01$, standard deviation 0.33, range -0.15 to 1).

Question 1: Impact of Familiarity upon Intelligibility Judgments

The relationship between the DME ratings for Sentence C compared to Sentence R from each talker is displayed in Figure 3-4 in order to allow for direct comparison of stimuli. Reliability for the direct magnitude estimation judgments was calculated with the

Log_{10} of the average raw perceptual scores because these judgments had been made on a ratio scale. The Pearson's correlation between the two sentences from each talker was r=0.82, p<0.01.

In order to more closely examine the impact of severity of dysarthria upon magnitude estimation, the 80 talkers were sorted by average listener rating (using the Log_{10} of rating averages) and organized into four severity groups (mild, mild-moderate, moderate, moderate-severe) of approximately 20 talkers. In figure 3-5 talkers were arranged from highest to lowest intelligible by the magnitude estimates across listeners. Sentence C ratings are generally set below the Sentence R curve.

Once arranged from least to greatest by the Log_{10} of the magnitude estimates across listeners, talkers were divided into the four groups. Pearson coefficients were calculated to compare Sentence C to Sentence R within each severity group. These

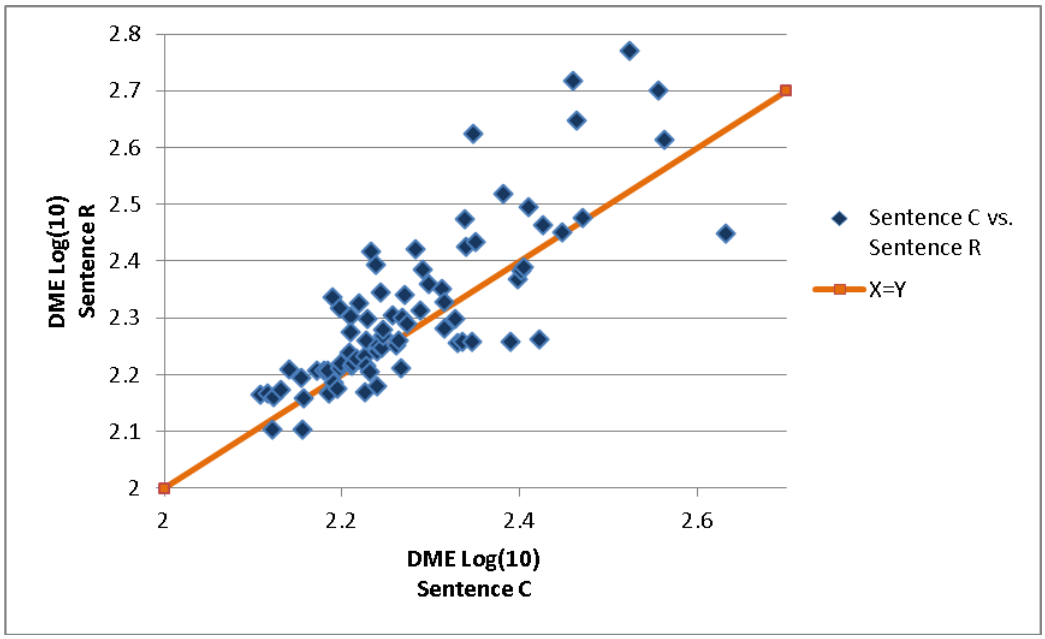

Figure 3-4. The Log_{10} conversion of absolute magnitude estimates for each talker averaged across listeners is plotted to compare Sentence C to Sentence R. (r = 0.82, p<0.01)

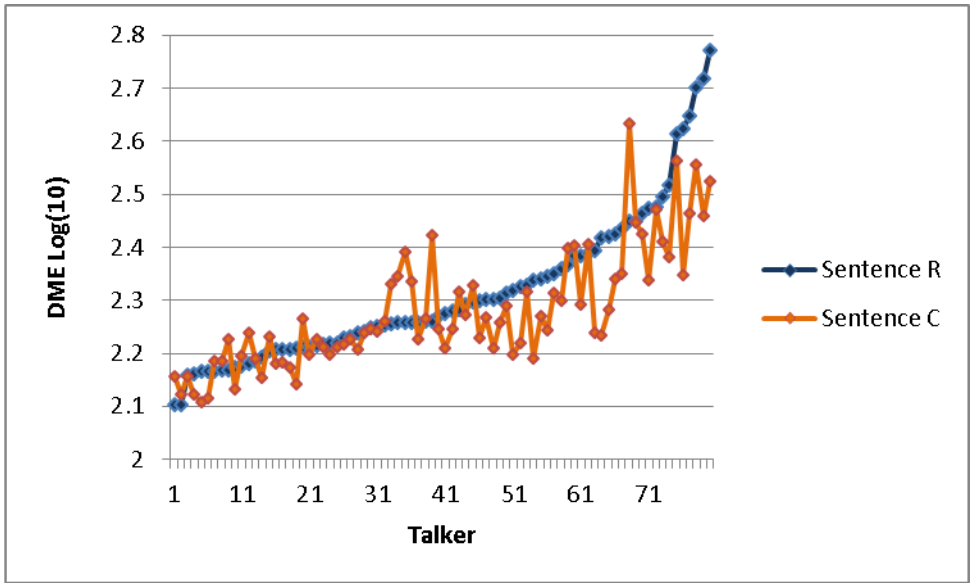

Figure 3-5. Talkers sorted by Sentence R ratings from highest to lowest intelligibility by listener judgments on the DME task.

Table 3-1. Correlations between Sentence C and Sentence R within each of the four severity ratings.

	Mild	Mild-Moderate	Moderate	Mod.-Severe
Group Total *n*	19	23	21	17
R	0.366804	0.475537	0.41945	0.499261
R^2	0.134545	0.226135	0.175938	0.249261

analyses are represented in Table 3-1 and reveal weaker correlations than when all 80 talkers are combined. This finding may be a function of the reduced number of data points and the smaller range of intelligibility variation within each category.

Question 2: Correlation of DME versus Transcription

A Pearson's correlation of the average rating from the DME task and the Transcription task revealed a strong inverse relationship ($r=-0.781$, $p<0.01$). The Log_{10} of the average DME and the arcsin of the decimal percentages (percentage/100) from the Transcription task were utilized in analysis. The DME data for each talker is plotted in relation to the Transcription data in Figure 3-6. Only the 30 least intelligible talkers

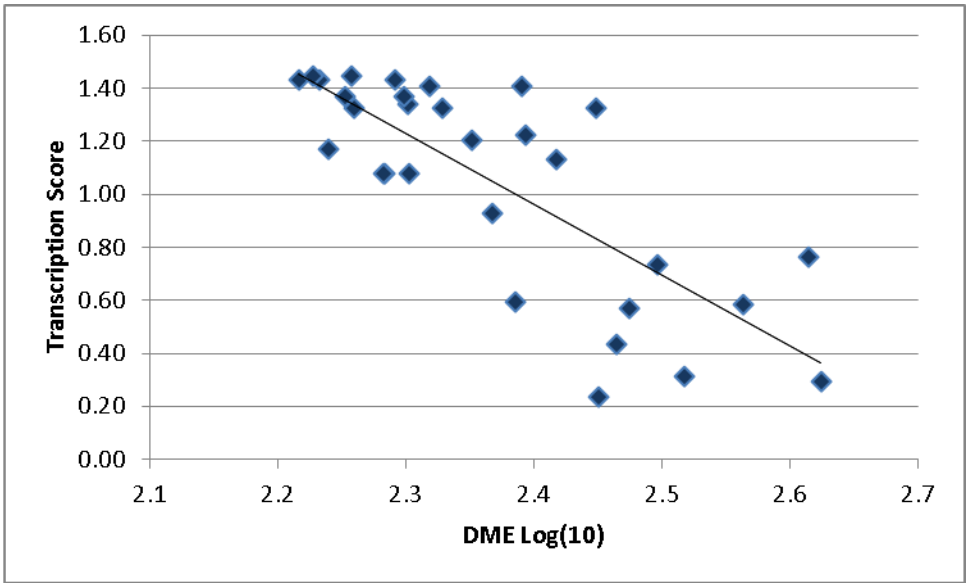

Figure 3-6. The Log_{10} of the average DME rating for each talker was plotted in relation to the arcsin of the decimal percentage of the transcription score ($r=-0.781$, $p<0.01$).

were analyzed, as the other half of the talkers were so highly intelligible that they all received 100% intelligibility scores on the Transcription task.

A Spearman's Rank Order correlation was run to determine the relationship between the rank order results from the DME task and those from transcription, again only correlating the 30 least intelligible talkers. The transcription data were sorted from greatest to least. The DME data were ordered opposingly, from lowest to highest, in order to rank the listener perceptions of the talkers in the appropriate direction, as high transcription scores denote high intelligibility whereas high ratings on the DME task signify poor intelligibility. Figure 3-7 depicts a significant positive correlation ($r_s=0.777$, $p<0.01$) between talker rank order derived from DME and transcription data. Higher ratings on the DME task corresponded to poorer transcription scores, indicating that both measures result in similar order rankings in regard to degree of intelligibility.

Figure 3-7. The rank order of each talker according to the results of both DME and Transcription are compared, with 1 denoting least intelligibility and 60 denoting highest intelligibility (r_s=0.777, $p<0.01$).

In order to more closely examine the impact of the two different intelligibility tasks upon perceptual judgments, the average talker data were organized first from poorest to highest intelligibility according to the DME rank order data. The DME data were then correlated with the Transcription data from the corresponding talkers. In the same way, talkers were then arranged from least to greatest according to the transcription data and then the Transcription data were correlated to corresponding DME data for talkers in this new rank ordering. This was done in order to group talkers according to perceived dysarthric severity by both intelligibility measures and to better assess whether discrepancies between the data from the two tasks had higher incongruity in the mild, moderate, or severe groupings.

Pearson's coefficients were found between the Log_{10} of the average talker ratings from the DME task and the arcsin of the decimal percentages from Transcription according to talker intelligibility. Table 3-2 displays the correlations between these two

Table 3-2. Pearson's coefficient found between Log_{10} of average DME and arcsin of decimal percentage of Transcription averages, arranged and correlated by severity (1 = poor intelligibility, 60 = high intelligibility).

Rank Ordered by DME			Rank Ordered by Transcription		
# Correlated	R	p-value	# Correlated	R	p-value
1-10	-0.319	>0.01	1-10	-0.251	>0.01
1-20	-0.695	<0.01	1-20	-0.684	<0.01
1-30	-0.743	<0.01	1-30	-0.757	<0.01
1-40	-0.764	<0.01	1-40	-0.758	<0.01
20-40	-0.07	>0.01	20-40	-0.348	>0.01
31-60	-0.300	<0.01	31-60	*	*

intelligibility measurement tools, arranged and correlated by severity. The rank ordering system implemented results in lower numbers denoting lower intelligibility and higher numbers signifying higher intelligibility. The first two groupings of 1-10 and 1-20 represent severe and moderate-severe groupings, respectively. The 20-40 and 31-60 groupings would correspond to mild-moderate and mild severity groupings, respectively. These correlation statistics reveal reduced strength at the extremes as well as in the isolated mild-moderate grouping (20-40). Correlations improve when half or two-thirds of the talkers are compared, from the more severely unintelligible extreme.

CHAPTER 4
DISCUSSION

The development of an automated measurement tool which could predict the average judgments of naïve listeners would be optimal for patient diagnostics and treatment. Ultimately this research seeks to develop automated measures of intelligibility using automatic speech recognition technology and signal distortion measures. Once developed, these algorithms will serve as an automated measurement tool which can predict the average listener judgments of speech intelligibility with novel stimuli.

A large database of speech samples from hundreds of people diagnosed with PD has been created. This database could be used in the development of automatic speech recognition systems and as an ideal baseline for test development.

Two primary experiments have been conducted. The first experiment included the training and extensive testing of listeners on a DME task as a measure of speech intelligibility of talkers diagnosed with PD. The main question of this experiment investigates the impact of stimuli used, and analyzes the impact of using one single stimulus versus multiple different stimuli in the context of making intelligibility judgments. The second experiment consisted of one-hundred listeners participating in a one-time, ten minute transcription task in which they were presented with just 15 different sentences from 15 different talkers. The main question of this second experiment was to correlate the DME data to the Transcription data, identifying differences between these two methods of intelligibility measurement. Additionally, this work will help create baseline perceptual data that may be used for the development of automatic measures of speech intelligibility in patients with dysarthria subsequent to PD.

Conclusions

Question 1: Impact of Familiarity upon Intelligibility Judgments

The Log_{10} of the averages across listeners of both Sentence C and Sentence R from each talker were correlated, revealing a strong positive correlation between both sentences ($r=0.82$, $p<0.01$). Figures 3-4 and 3-5 demonstrate that, while both sentences are highly correlated, there is a clear tendency for the listeners to rate Sentence R as more highly unintelligible. In Figure 3-4, the majority of data points are above the x=y line, expressing that DME judgments for Sentence R generally exceeded those of Sentence C. Likewise, Figure 3-5 displays the Sentence C curve as generally below the Sentence R curve, reiterating the fact that Sentence C was consistently judged to be more highly intelligible than Sentence R. Such a trend may be a function of high familiarity with Sentence C, recall this sentence was presented 5 times for each of the 80 talkers in addition to the standard which was reiterated 45 times throughout the experiment (total = 445), compared to the mere 20-35 times a listener would rate one of the other 14 sentences. Additionally, the same comparison revealed tighter data for talkers with better perceived intelligibility and the data progressively increased in variability as perceived magnitude of unintelligibility ascends. In other words, it appears that the intelligibility judgments of more severely dysarthric talkers are more difficult to predict.

Question 2: Correlation of DME versus Transcription

The correlation of the DME and Transcription tasks revealed a significant inverse relationship ($r=-0.781$, $p<0.01$). Because a low score on the DME task denotes high intelligibility and a high score on the Transcription task also signifies high intelligibility,

the inverse nature of the relationship between these two tasks is represented by a negative r value.

The Spearman's rank order correlation between both intelligibility measures revealed a significant positive correlation (r_s=0.777, p<0.001) between talker rank order. That is, higher ratings on the DME task corresponded to poorer percent intelligibility scores, indicating that both measures result in similar order rankings of degree of intelligibility. While statistically significant, the strength of this correlation is lacking.

In order to more closely examine the nuanced correlation of the two different intelligibility tasks, the average talker data were arranged and correlated by severity, as referenced in Table 3-2. The rank ordering system was characterized by lower numbers denoting lower intelligibility and higher numbers signifying higher intelligibility. The correlation between the data from the DME and transcription tasks was strongest across a large number of talkers. When particular severity groupings were targeted in order to analyze the correlation between data from the two tasks, the Pearson’s coefficient reduced dramatically. For example, correlations between severity groups including 1-10, 20-40, and 31-60, in both rank order scenarios, each resulted in relationships which lacked the strength to be statistically significant, even though the mild-moderate and mild groupings, 20-40 and 31-60 respectively, exceeded an *n* of 15.

While there is a positive correlation between the two intelligibility tasks, there is higher variability between the tasks in talkers who are moderately unintelligible. Due to high agreement between listener perceptions at the extremes of the scale, it appears that the task of rating someone with high intelligibility or extremely low intelligibility is

more straightforward than defining the wide moderate range. Listener agreement with talkers in the middle of the intelligibility spectrum is more of a challenge.

Numerous mild and mild-moderate talkers were represented in this study, resulting in 30 average transcription scores equaling 100% accuracy; therefore correlations could not be made across the 30 mildest talkers when rank-ordered by transcription score, due to the fact that the transcription variable was constant in this grouping. Correlations are strongest across both intelligibility measures when the less intelligible half or two-thirds is compared, revealing a strong negative correlation between these data. Due to reduced variability in the highly intelligible data, especially from the transcription task, and consequent decline in correlation strength, it becomes less meaningful to compare this group, as there appears to be higher discrepancy between the data across both tasks.

APPENDIX A
STIMULI

Sentences

1. His boss made him work like a slave
2. He caught the fish in his net
3. The beer drinkers raised their mugs
4. I made the phone call from a booth
5. The cut on his knee formed a scab
6. I gave her a kiss and a hug
7. The soup was served in a bowl
8. The cookies were kept in a jar
9. The baby slept in his crib
10. The cop wore a bullet proof vest
11. How long can you hold your breath?
12. At breakfast he drank some juice
13. I ate a piece of chocolate fudge
14. The judge is sitting on the bench
15. The boat sailed along the coast

The Rainbow Passage

When the sunlight strikes raindrops in the air, they act as a prism and form a rainbow. The rainbow is a division of white light into many beautiful colors. These take the shape of a long round arch, with its path high above and its two ends apparently beyond the horizon. There is, according to legend, a boiling pot of gold at the end. People look, but no one ever finds it. 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.

APPENDIX B
SUBJECT INFORMATION FORM

Subject Code:_____

Date_____

SUBJECT INFORMATION FORM

Name:_____
Age (in years):_____**Birthdate:**_____

Sex: M F

Ethnicity: _____

1) **First language spoken if not English (or in addition to English):**_____

2) **List any other languages you speak, along with your proficiency:**_____
_____
_____

3) **Have you ever had a hearing or speech disorder?** Yes No
If "Yes," please explain:_____

4) **Major and Year** _____

APPENDIX C
GRADING TEMPLATE

1 his BOSS(1) MADE(1) him WORK(1) like a SLAVE(1) = 4
2 he CAUGHT(1) the FISH(1) in his NET(1) = 3
3 the BEER(1) DRINK(1)ERS(1) RAISED(1) their MUGS(1) = 5
4 i MADE(1) the PHONE(1) CALL(1) from a BOOTH(1) = 4
5 the CUT(1) on his KNEE(1) FORMED(1) a SCAB(1) = 4
6 i GAVE(1) her a KISS(1) and a HUG(1) = 3
7 the SOUP(1) was SERVED(1) in a BOWL(1) = 3
8 the COO(1)KIES(1) were KEPT(1) in a JAR(1) = 4
9 the BA(1)BY(1) SLEPT(1) in his CRIB(1) = 4
10 the COP(1) WORE(1) a BUL(1)LET(1) PROOF(1) VEST (1) = 6
11 how LONG(1) can you HOLD(1) your BREATH(1) = 3
12 at BREAK(1)FAST(1) he DRANK(1) some JUICE(1) = 4
13 i ATE(1) a PIECE(1) of CHO(1)COLATE(1) FUDGE(1) = 5
14 the JUDGE(1) is SIT(1)TING(1) on the BENCH(1) = 4
15 the BOAT(1) SAILED(1) A(1)LONG(1) the COAST(1) = 5

LIST OF REFERENCES

Beukelman, D.R., & Yorkston (1979). The relationship between information transfer and speech intelligibility of dysarthric speakers. *Journal of Communication Disorders 12*(3), 189-196.

Beukelman, D.R., & Yorkston K.M. (1980). Influence of passage familiarity on intelligibility estimates in dysarthric speech. *Journal of Communication Disorders, 13*(1), 33-41.

Darley, F. L., Aronson, A. E., & Brown, J. R. (1969a). Clusters of deviant speech dimensions in the dysarthrias. *Journal of Speech and Hearing Research, 12*(3), 462-496.

Darley, F. L., Aronson, A. E., & Brown, J. R. (1969b). Differential diagnostic patterns of dysarthria. *Journal of Speech and Hearing Research, 12*(2), 246-269.

De Bodt, M. S., Hernandez-Diaz, H. M., & Van De Heyning, P. H. (2002). Intelligibility as a linear combination of dimensions in dysarthric speech. *Journal of Communication Disorders, 35*(3), 283-292.

Di Benedetto, P., Cavazzon, M., Mondolo, F., Rugiu, G., Peratoner, A., & Biasutti, E. (2009). Voice and choral singing treatment: a new approach for speech and voice disorders in Parkinson's disease. *European Journal of Physical and Rehabilitation Medicine, 45*(1), 13-19.

Duffy, J.R. (2005). Motor Speech Disorders: substrates, differential diagnosis, and management. Elsevier Health Sciences.

Fairbanks, G. (1960). Voice and articulation drill book, 2nd edition. New York: Harper & Row. p.124-139.

Fischer, E., & Goberman, A.M. (2010). Voice onset time in Parkinson disease. *Journal of Communication Disorders, 43*(1) 21-34.

Frost, E., Tripoliti, E., Hariz, M.I., Pring, T., & Limousin, P. (2010). Self-perception of speech changes in patients with Parkinson's disease following deep brain stimulation of the subthalamic nucleus. *International Journal of Speech Language Pathology,12*(5), 399- 404.

Goberman, A. M., & Coelho, C. (2002a). Acoustic analysis of parkinsonian speech I: speech characteristics and L-Dopa therapy. *NeuroRehabilitation, 17*(3), 237-246.

Goberman, A. M., & Coelho, C. (2002b). Acoustic analysis of parkinsonian speech II: L-Dopa related fluctuations and methodological issues. *NeuroRehabilitation, 17*(3), 247-254.

Ho, A.K., Iansek, R., & Bradshaw, J.L. (2001). Motor instability in parkinsonian speech intensity. *Neuropsychiatry, Neuropsychology & Behavioral Neurology. 14*(2), 109-16.

Hoehn, M., & Yahr, M. (1967). Parkinsonism: Onset, progression, and mortality. *Neurology, 17*, 427-442.

Kalikow, D.N., Stevens, K.N., & Elliott, L.L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *Journal of the Acoustical Society of America*, 61, 1337-1351.

Kent, R. D., Weismer, G., Kent, J. F., & Rosenbek, J. C. (1989). Toward phonetic intelligibility testing in dysarthria. *J Speech Hear Disord, 54*(4), 482-499.

Kent, R. D., & Kent, J. F. (2000). Task-based profiles of the dysarthrias. *Folia Phoniatr Logop,* 52, 48-53.

Kreiman, J., Gerratt, B.R., Kempster, G.B., Erman, A., & Berke, G.S.. (1993). Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research*. Journal of Speech, Language, & Hearing Research,* 36, 21–40.

Oguz, H., Tunc, T., Safak, M.A., Inan, L., Kargin, S., & Demirci, M. (2006). Objective voice changes in nondysphonic Parkinson's disease patients. *Journal of Otolaryngology, 35*(5), 349-354.

Pahwa, R., & Lyons, K.E. (2010). Early diagnosis of Parkinson's disease: recommendations from diagnostic clinical guidelines. *American Journal of Managed Care. 16*(3), 94-99.

Quedas, A., Duprat Ade, C., & Gasparini, G. (2007). Lombard's effect's implication in intensity, fundamental frequency and stability on the voice of individuals with Parkinson's disease. *Brazian Journal of Otorhinolaryngology, 73*(5), 675-683.

Sakar, C.O., & Kursun, O. (2010). Telediagnosis of Parkinson's disease using measurements of dysphonia. *Journal of Medical Systems, 34*(4), 591-599.

Sarno, M. T, (1968). Speech impairment in Parkinson’s disease. *Archives of Physical Medicine and Rehabilitation,* 49, 269-275.

Sapienza, C., Troche, M., Pitts, T., & Davenport, P. (2011). Respiratory strength training: concept and intervention outcomes. *Seminars in Speech & Language, 32*(1), 21-30.

Stevens, S.S., Volkmann, J., & Newman, E.B. (1937). A Scale for the Measurement of the Psychological Magnitude Pitch. *Journal of the Acoustical Society of America, 8*(3), 208-208.

Stevens, S. S. (1946). On the Theory of Scales of Measurement. Science, 103(2684), 677-680.

Weismer, G., & Martin, R. E. (1992). Acoustic and perceptual approaches to the study of intelligibility. In R. D. Kent (Ed.), *Intelligibility in speech disorders: Theory, measurement and management*. Amsterdam: John Benjamins.

Weismer, G., Jeng, J. Y., Laures, J. S., Kent, R. D., & Kent, J. F. (2001). Acoustic and intelligibility characteristics of sentence production in neurogenic speech disorders. *Folia Phoniatrica Logopaedica, 53*(1), 1-18.

Weismer, G., & Loures, J. S. (2002). Direct magnitude estimates of intelligibility in dysarthria: effects of a chosen standard. *Journal of Speech, Language, and Hearing Research, 45*(3), 421-433.

Yorkston, K. M., & Beukelman, D. R. (1978). A comparison of techniques for measuring intelligibility of dysarthric speech. *Journal of Communicative Disorders, 11*(6), 499-512.

Yorkston, K. M., & Beukelman, D. R. (1980). A clinician-judged technique for quantifying dysarthric speech based on single-word intelligibility. *Journal of Communicative Disorders, 13*(1), 15-31.

Yorkston, K. M., Strand, E. A., & Kennedy, M. R. T. (1996). Comprehensibility of dysarthric speech: Implications for assessment and treatment planning. *American Journal of Speech-Language Pathology, 5*(1), 55-66.

Zarzur, A.P., Duprat, A.C., Shinzato, G., & Eckley, C.A. (2007). Laryngeal electromyography in adults with Parkinson's disease and voice complaints. *Laryngoscope, 117*(5), 831-834.

BIOGRAPHICAL SKETCH

Traci Reynolds earned her bachelor's degree from The University of Florida, where she majored in communication sciences and disorders with minors in religion as well as international development and humanitarian assistance. She recently completed a Master's of Arts in speech-language pathology at the University of Florida and is excited to transition into full-time clinical work. She greatly looks forward to a life of service to the local and global community!