THE EFFECT OF EXPERIENCE AND THE RELATIONSHIP AMONG
SUBJECTIVE AND OBJECTIVE MEASURES OF VOICE QUALITY

A Dissertation in
Communication Sciences and Disorders
by
Jessica L. Sofranko

© 2012 Jessica L. Sofranko

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2012
The dissertation of Jessica L. Sofranko was reviewed and approved* by the following:

Robert A. Prosek
Professor of Communication Sciences and Disorders
Dissertation Adviser
Chair of Committee

Norman Spivey
Professor of Music

Kathryn D. R. Drager
Associate Professor of Communication Sciences and Disorders
Program Director

Gordon W. Blood
Professor of Communication Sciences and Disorders
Head of the Department of Communication Sciences and Disorders

*Signatures are on file in the Graduate School.
ABSTRACT

The purpose of this study was to determine the effect of experience on the judgment of voice quality, to determine the perceptual space being used across these groups when judging voice quality, and to examine the correlation among acoustical measurements and perceptions of voice quality. Speech-language pathologists, singing voice teachers, speech-language pathology graduate students with and without experience with a voice client, graduate students who have completed a voice pedagogy course, and inexperienced listeners participated in two experiments. In Experiment One, participants rated stimuli with systematically altered measurements of jitter, shimmer, and noise-to-harmonics ratio (NHR) ranging from mild to severe for overall severity, roughness, breathiness, strain, and pitch. In Experiment Two, participants rated the similarity of the same stimuli in pairs. Results from Experiment One showed that type of experience had an impact on judgments of voice quality more than level of experience. Also, jitter/shimmer combination stimuli and shimmer only stimuli frequently correlated with ratings of overall severity, roughness, and strain and NHR stimuli correlated with ratings of breathiness across all groups. Only inexperienced listeners, singing voice teachers, and their students had significant correlations for ratings of pitch with jitter/shimmer combination stimuli having the highest correlations. Results from Experiment Two showed that participants with different levels and types of experience used different perceptual spaces (of additive noise and perturbation measures) when judging similarity of stimulus pairs. The conclusion was that level and type of experience affects judgments of voice quality.

Keywords: voice perception, experienced listener, multidimensional scaling, listener agreement, acoustical measures
# TABLE OF CONTENTS

List of Tables .......................................................................................................................... vi
List of Figures ............................................................................................................................ vii
PREFACE .................................................................................................................................. viii

CHAPTER 1. INTRODUCTION ........................................................................................................ 1
  1.1. Internal and External Standards .................................................................................. 4
      1.1.1. Experience ........................................................................................................... 5
  1.2. Rating Scale ..................................................................................................................... 8
      1.2.1. Multidimensional Scaling .................................................................................... 9
  1.3. Acoustical Measures ....................................................................................................... 11
  1.4. Stimulus Factors ........................................................................................................... 12
      1.4.1. Synthesized Stimuli ............................................................................................ 13
      1.4.2. Stimulus Length ................................................................................................... 13
  1.5. Summary ......................................................................................................................... 14
      1.5.1. Interrater Agreement ........................................................................................... 14
      1.5.2. Standardized Rating Scales .................................................................................. 14
      1.5.3. Perceptions of Voice Quality ............................................................................... 16
      1.5.4. Précis .................................................................................................................... 16
  1.6. Purpose ........................................................................................................................... 17
  1.7. Research Questions ....................................................................................................... 17

CHAPTER 2. METHODS ................................................................................................................. 19
  2.1. Experiment One ............................................................................................................. 19
      2.1.1. Stimuli .................................................................................................................. 19
      2.1.2. Listeners .............................................................................................................. 22
      2.1.3. Procedures .......................................................................................................... 25
  2.2. Experiment Two ............................................................................................................. 27
      2.2.1. Stimuli .................................................................................................................. 27
      2.2.2. Listeners .............................................................................................................. 27
      2.2.3. Procedures .......................................................................................................... 27

CHAPTER 3. RESULTS ..................................................................................................................... 29
  3.1. Experiment One ............................................................................................................. 29
      3.1.1. Intrarater Agreement ........................................................................................... 29
      3.1.2. Interrater Agreement ........................................................................................... 30
      3.1.3. Correlation .......................................................................................................... 31
      3.1.4. Scale Analysis ...................................................................................................... 36
          3.1.4.1. Overall Severity ........................................................................................... 37
          3.1.4.2. Roughness ................................................................................................... 39
          3.1.4.3. Breathiness .................................................................................................. 41
          3.1.4.4. Strain .......................................................................................................... 42
          3.1.4.5. Pitch ............................................................................................................. 44
      3.1.5. Effect Size ............................................................................................................. 46
  3.2. Experiment Two .............................................................................................................. 48
List of Tables

Table 3.1  Summary of Intrarater Agreement ...............................................................30
Table 3.2  Summary of Interrater Agreement ...............................................................31
Table 3.3  Group: Tukey HSD Results for Overall Severity ........................................37
Table 3.4  Type of Stimulus: Tukey HSD Results for Overall Severity .......................38
Table 3.5  Group: Tukey HSD Results for Roughness ..................................................39
Table 3.6  Type of Stimulus: Tukey HSD Results for Roughness ...............................40
Table 3.7  Group: Tukey HSD Results for Breathiness ................................................41
Table 3.8  Type of Stimulus: Tukey HSD Results for Breathiness ...............................42
Table 3.9  Group: Tukey HSD Results for Strain .......................................................43
Table 3.10 Type of Stimulus: Tukey HSD Results for Strain ......................................44
Table 3.11 Group: Tukey HSD Results for Pitch .........................................................45
Table 3.12 Type of Stimulus: Tukey HSD Results for Pitch ......................................45
Table 3.13 Effect Size: MANOVA Results .................................................................47
Table 3.14 Effect Size: ANOVA Results ..................................................................47
Table 3.15 S-Stress Improvement ..............................................................................49
Table 3.16 Group Stress Levels and R² Values .........................................................49
Table 3.17 Group Matrices Weirdness Values and Weights .......................................50
### List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Visual analog screen display in Alvin2</td>
<td>26</td>
</tr>
<tr>
<td>2.2</td>
<td>Multidimensional scaling screen layout in Alvin2</td>
<td>28</td>
</tr>
<tr>
<td>3.1</td>
<td>Correlations between Types of Stimuli and Overall Severity</td>
<td>32</td>
</tr>
<tr>
<td>3.2</td>
<td>Correlations between Types of Stimuli and Roughness</td>
<td>33</td>
</tr>
<tr>
<td>3.3</td>
<td>Correlations between Types of Stimuli and Breathiness</td>
<td>34</td>
</tr>
<tr>
<td>3.4</td>
<td>Correlations between Types of Stimuli and Strain</td>
<td>35</td>
</tr>
<tr>
<td>3.5</td>
<td>Correlations between Types of Stimuli and Pitch</td>
<td>36</td>
</tr>
<tr>
<td>3.6</td>
<td>Mean Ratings for Overall Severity</td>
<td>38</td>
</tr>
<tr>
<td>3.7</td>
<td>Mean Ratings of Roughness</td>
<td>40</td>
</tr>
<tr>
<td>3.8</td>
<td>Mean Ratings for Breathiness</td>
<td>42</td>
</tr>
<tr>
<td>3.9</td>
<td>Mean Ratings for Strain</td>
<td>43</td>
</tr>
<tr>
<td>3.10</td>
<td>Mean Ratings for Pitch</td>
<td>45</td>
</tr>
<tr>
<td>3.11</td>
<td>Scree Plot</td>
<td>49</td>
</tr>
<tr>
<td>3.12</td>
<td>Groups Plotted as D1 Vs. D2</td>
<td>52</td>
</tr>
<tr>
<td>3.13</td>
<td>Groups Plotted as D2 Vs. D3</td>
<td>53</td>
</tr>
<tr>
<td>3.14</td>
<td>3D Plot of Group Weights</td>
<td>54</td>
</tr>
</tbody>
</table>
PREFACE

Dr. Prosek, where do I begin? I chose Penn State because of you. Over these past three years you have shared your knowledge, expertise, advice, and humor. You have taught me not only speech science, but the ins and outs of teaching, presenting, defending, and dealing with struggles that I didn’t anticipate. I can’t believe the countless times you have agreed to read my work, frequently saying, “I’ll just wait and catch the movie.” You were so patient to be sure I learned principles that will benefit me throughout my career. You have often described what it was like to work with me and now it’s my turn. Working with you has been an honor. I am so incredibly lucky to have had the opportunity to learn from the best. Without you, none of this would have been possible and I can’t even begin to thank you for all of your guidance and support. I am proud to call you my mentor and my friend. I would not be where I am today if it wasn’t for you.

I would also like to thank the members of my committee. Dr. Blood, I need to thank you for making me a strong, competent, and determined researcher and teacher. Because of you, I have learned to refine my goals, know my boundaries, and strive to be the best. Dr. Drager, your guidance through this process provided me with a backbone which ultimately brought me to my goal. You have also given me opportunities to become a better leader and make a difference. Also, thank you for inviting me to teach in your classroom and share another one of my passions. Your time, patience, and knowledge are greatly appreciated. Dr. Spivey, thank you for sharing your gift and reminding me why I became interested in voice in the first place. Your passion for voice and music is an inspiration and I was so incredibly lucky to have you as a teacher. Your generosity and welcoming invitation into your department allowed me to expand my knowledge
in the area of voice in a way I couldn’t imagine. You made me a better researcher, a better teacher, a better singer, and a better person.

Thank you, Drs. Cari and Glen Tellis, for inspiring me and motivating me to obtain my Ph.D. Your dedication to your work and students has been an example I hope to live up to as I start my career. Glen, you have been my guide throughout every process and have taught me what it is to be a true professional. You have given me countless opportunities to grow and learn in this field and I will be forever grateful. Cari, because of you, I had the confidence to face challenges throughout this process and I have the confidence to face challenges to come in the future. You taught me to focus on what’s important and trust in myself. I would be lucky to be half the teacher and researcher you are in the field. I thank you both for being mentors and friends. You have done so much for me over the years. This is for both of you.

Lastly, none of this would have been possible without the love and support from my family and friends. Thank you all! Mom, dad, you taught me to strive to be the best and maintain the humility needed to succeed. You believed in me from the very start and your unconditional support provided me with the confidence and the drive to continue to reach my goals. You were both there for the good times and the bad and this is as much for you as it is for me. Mike, thank you for providing me with the perspective of what is really important. You came into my life while I was in the middle of this process, but you made a stressful time the best time of my life. You were patient with my hectic schedule, understanding of my frequent stresses, and most importantly, supportive every step of the way. Continuously selfless, you were there for some of the most difficult times. You listened to endless conversations ultimately making you quite the voice expert! I couldn’t have done it without you. Thank you.
1. CHAPTER 1 - INTRODUCTION

“Acoustical communication is one of the fundamental prerequisites for the existence of human society” (Fastl & Zwicker, 2007, p. VII). In order to accomplish this acoustical communication, our auditory system processes both qualitative information, one’s impression of the signal, and quantitative information, one’s sensory input from the signal (Fastl & Zwicker, 2007). The study of the relationship between the sounds an individual perceives and his or her experiences, including both objective and subjective descriptions of sound quality, is known as psychoacoustics (Cook, 1999; Parn cott, 2004). Cognitive or aesthetic effects can be present during evaluations of sound quality (Fastl & Zwicker, 2007). This interaction between the sound and one’s impression includes judgments based on semantic, syntactic, environmental, and vocal information when determining communication intent (Moore, 1989). These judgments are shaped by one’s experience, attention, and memory (Gerratt, Kreiman, Antonanzas-Barroso, & Berke, 1993; Kreiman, Gerratt, Precoda, & Berke, 1992).

Psychoacoustics applies to many fields as it is relevant to both sounds we receive and sounds transmitted through speech (Fastl & Zwicker, 2007). Sound quality, in this case specifically voice quality, is a subjective measure made by a listener. Perception of communication is not solely based on the signal itself, but also by the knowledge of the listener (Lindblom, 1990). Some subjective measures of voice quality in the field of speech-language pathology include: pitch, loudness, strain, breathiness, and roughness. Singing voice teachers may be more inclined to focus on more aesthetic qualities of voice such as the amount of: ring, focus, balance, and vowel color (Miller, 1996). When studying sound quality, listeners judge what they hear based upon several attributes of interest specific to them (Fastl, 2005). Several decades of research in the area of voice perception show that not only level and type of
experience affect and form a listener’s attributes of interest, but also the presence of anchors, the type of rating scale, the type of stimuli, and the length of stimuli (Bassich & Ludlow, 1986; de Bruijn & Whiteside, 2007; de Krom, 1994; Eadie & Doyle, 2002; Fitch, 1990; Gerratt & Kreiman, 2001; Gerratt et al., 1993; Hillenbrand, 1987; Hillenbrand, 1988; Hillenbrand & Houde, 1996; Lee, Drinnan, & Carding, 2005; Kreiman & Gerratt, 1998; Kreiman & Gerratt, 2000; Kreiman, Gerratt, & Ito, 2007; Kreiman, Gerratt, Kempster, Erman, & Berke, 1993; Kreiman, Gerratt, & Precoda, 1990; Parsa & Jamieson, 2001; Rabinov, Kreiman, Gerratt, & Bielamowicz, 1995; Schoentgen, 1989; Shrivastav, Sapienza, & Nandur, 2005; Yiu, Chan, & Mok, 2007; Yiu & NG, 2004; Zraick & Liss, 2000). In fact, Kreiman et al., (2007) found that varying internal standards (including level and type of experience), difficulty isolating targeted attribute(s) in a multidimensional signal, the type of scale being used to rate voice quality, and the degree of the characteristic that is being measured accounted for 84.2% of the variance for listener agreement.

Throughout the literature, many attempts have been made to examine these factors and their effects on listener perception; however, research has been contradictory. It can be argued that one reason for the contradictions may be due to a lack of experimental control across all identified variables affecting listener perception when examining each factor individually. The effect of experience on listener perceptions of voice quality needs to be determined for proper group selection prior to experimentation. All known variables that affect perceptions of voice quality need to be controlled to determine if the differences in level and type of experience have an impact on listener judgments. After the effect of experience is known, then researchers can appropriately choose listeners for studies aimed to examine relationships among acoustical measures and perceptions of voice quality.
The aforementioned variables, (e.g. internal standards, type of rating scale, and type and length of stimulus) can affect perceptual judgments of voice quality, and in turn affect correlations with acoustical measurements of voice (Bassich & Ludlow, 1986; de Bruijn & Whiteside, 2007; de Krom, 1994; Eadie & Doyle, 2002; Fitch, 1990; Gerratt & Kreiman, 2001; Gerratt et al., 1993; Hillenbrand, 1987; Hillenbrand, 1988; Hillenbrand & Houde, 1996; Lee et al., 2005; Kreiman & Gerratt, 1998; Kreiman & Gerratt, 2000; Kreiman et al., 2007; Kreiman et al., 1993; Kreiman et al., 1990; Parsa & Jamieson, 2001; Rabinov et al., 1995; Schoentgen, 1989; Shrivastav et al., 2005; Yiu et al., 2007; Yiu & NG, 2004; Zraick & Liss, 2000). In addition, because the research methodology varies, understanding of perceptions of voice quality is likely limited. Most importantly, there are very few studies examining the differences between experienced and inexperienced listeners for perceptions of voice quality (Kreiman et al., 1990).

In spite of this, generalizations are made, not only for relationships among acoustical measurements and perceptions, but for appropriate rating scales and types and length of stimuli using speech-language pathology graduate students as judges. However, speech-language pathology graduate students may not be considered expert or experienced listeners as these students have just started their clinical career and still require supervision. Also, they do not fit the operational definitions that some researchers have been using to define an expert listener throughout the literature.

The effects of experience on perceptions have been studied in many formats including infant speech perception, listeners’ perception of different dialects, and listeners’ discrimination of their native language (Kreiman, Gerratt, & Khan, 2010; Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992; Sumner & Samuel, 2009; Werker & Tees, 1984). Results have shown that language stimulation, or experience, leads to a “perceptual system that highlights the contrasts
used in language, while de-emphasizing those that do not” (Gazzaniga, 2000, p. 103). Experience also has been shown to affect a listener’s ability to recognize spoken words in a specific dialect (Sumner & Samuel, 2009). Native language can also affect a listener’s sensitivity to acoustical variations within the signal (Kreiman et al., 2010). Therefore, continued experience, be it listening to voices on a daily basis and/or within academic coursework in the area of voice/voice disorders, should affect one’s internal standards, and ultimately, perceptions when rating voice quality. This introduction will begin reviewing the known factors that affect voice quality perceptions beginning with internal standards followed by types of rating scales, relationships with acoustical measures, and stimulus factors.

1.1. Internal and External Standards

Internal standards, used for perceptual ratings, are defined as an individual’s baseline for judgment which is affected by his or her experience, memory, and/or attention (Kreimen, Gerratt, & Berke, 1994; Kreiman et al., 1993). Also known as individual differences, internal standards are unstable and can lead to disagreement among listeners (Kreiman, Vanlancker-Sidtis, & Gerratt, 2004). Much like language experience, a difference in voice experience can lead to a specific sensitivity or bias (Eadie & Baylor, 2006).

The effect of internal standards on perceptions of voice quality is well documented. Research studies have successfully minimized the effect of internal standards through the use of anchors (external standards). By using fixed external standards, such as a synthesized representation of normal voice quality, results indicate an increase in interrater and intrarater agreement (Bassich & Ludlow, 1986; Gerratt et al., 1993; Kreiman et al., 2007; Yiu et al., 2007). These anchors help the listener focus on only the difference between the stimulus and the anchor (Kreiman et al., 2007). An anchor provides each listener with the same baseline, attempting to
decrease the effect of his or her own personal experience. Only in the absence of a fixed external standard, is one able to determine the effects of level and type of experience on perceptions of voice quality.

1.1.1. Experience

Experience is another factor that may affect perceptions of voice quality. Expert listeners are said to have access to a larger collection of information related to voice quality perceptions when listening to and judging voice quality as compared to inexperienced listeners with a varying educational background and a lack of consistent experience with dysphonia (Kreiman et al., 1990). Researchers typically create their own operational definitions of an expert or “experienced listener” ranging from job title only (e.g. speech-language pathologist, phonetician) to someone with 2 years of experience in the area of dysphonia (Anders, Hollien, Hurme, Sonnin, & Wendler, 1988; Kreiman et al., 1992). These experienced listeners, or generally those who are thought to have extensive background in the area of voice/voice disorders, not only use different components within the acoustical signal as compared to inexperienced listeners (Kreiman et al., 1992) but utilize additional resources within the signal (Kreiman & Gerratt, 1998; Kreiman et al., 1992). Ratings for normal voice quality, judged by experienced listeners (speech-language pathologists and otolaryngologists with a minimum of two years of experience), correlated with fundamental frequency, shimmer, and formant frequencies and only fundamental frequency correlated with judgments made by inexperienced listeners (Kreiman et al., 1992). Dysphonic voice quality judgments for experienced listeners correlated with fundamental frequency, partial period comparison, and the difference in amplitude between the first and second harmonics with inexperienced listeners’ dysphonic voice quality judgments correlating with fundamental frequency, the first formant frequency, the difference in amplitude
between the first and second harmonic, and perturbation measures (Kreiman et al., 1992). Also, clinicians, or experienced listeners, have been found to obtain a kappa statistic, or agreement, of 0.62 as compared to patients, or inexperienced listeners, rating their own voice quality with a kappa statistic of 0.32 (Lee et al., 2005).

Contrary to predictions from the literature regarding differences between experienced and inexperienced listeners, Anders et al. (1988) conducted a study comparing experienced and inexperienced listeners’ classifications of hoarse and normal voice quality and found no significant differences in voice quality perceptions among groups. Anders and colleagues defined their experienced listeners merely by job title. Speech-language pathologists, although given academic coursework in voice, may not treat voice disorders during their everyday practice. Also, their definition of experienced listeners included other types of backgrounds in addition to speech-language pathologists such as phoneticians and singing voice teachers. Additionally, some large differences of percentage correct when identifying normal and hoarse voices were averaged and experienced and inexperienced listeners appeared to perform similarly. Lastly, listeners were only given two choices and selection accuracy was compared to the classifications of the authors.

Patel, Shrivastav, and Cummings (2008) also compared the judgments of experienced and inexperienced listeners. The listeners were asked to add white noise (a band of frequencies with equal energy at each frequency often used to simulate breathiness) to a sawtooth wave (a complex wave with harmonics with both odd and even integer multiples, similar to the glottal output) for comparison to a synthesized breathy voice sample. Although the authors defined expert listeners as speech-language pathologists with experience in voice disorders, the amount of experience was not reported. Again, authors found no significant differences between
experienced and inexperienced listeners. Based on the methodology, the task used by Patel et al. (2008) may be considered a psychoacoustic matching task versus a voice perception task (Sofranko & Prosek, 2012).

Results of a recent study indicated that when defining groups using specific definitions of experience, there were significant differences in agreement between experienced and inexperienced listeners (Sofranko & Prosek, 2012). Experienced listeners included speech-language pathologists and singing voice teachers. Speech-language pathologists were individuals with over three years of experience with voice, spending 10 hours or more in the area of voice each week and singing voice teachers were individuals who were full members of the National Association of Teachers of Singing (NATS). In the past, experienced listeners have been defined as individuals with two or more years of experience with dysphonia (Rabinov et al., 1995; Kreiman et al, 1992), or individuals who have spent more than one third of his or her career over a three year period in the area of voice (De Bodt, Wuyts, Van de Heyning, and Croux, 1997) or individuals holding a Certificate of Clinical Competence (Patel et al., 2008). Inexperienced listeners were individuals from various backgrounds, with no previous experience in the area of voice including singing training and/or previous voice treatment. Overall, interrater agreement was significantly better for the experienced listener groups as compared to the inexperienced listeners.

A majority of the studies examining perceptions of voice quality have used only speech-language pathology graduate students as judges, listener groups containing both speech-language pathology graduate students and more experienced listeners such as speech-language pathologists experienced with dysphonia and/or otolaryngologists, or experienced listeners with different types of experience such as otolaryngologists, phoneticians, and speech-language
pathologists (Anders et al., 1988; Awan & Roy, 2006; Bassich & Ludlow, 1986; de Krom, 1994; Eadie & Doyle, 2002; Gerratt & Kreiman, 2001; Gerratt et al., 1993; Pausewang Gelfer, 1988; Gottlieb, Chew, Eshbaugh, & Prosek, 2010; Hillenbrand, Cleveland, & Erickson, 1994; Hillenbrand & Houde, 1996; Kempster, Kistler, & Hillenbrand, 1991; Kreiman & Gerratt, 1996; Kreiman, Gerratt, & Berke, 1994; Kreiman et al., 2007; Kreiman et al., 1993; Kreiman et al., 1990; Kreiman et al., 1992; Martens, Versnel, & Dejonckere, 2007; Murry, Singh, & Sargent, 1977; Rabinov et al., 1995; Shrivastav, 2006; Shrivastav et al., 2005; Smith, Weinberg, Feth, & Horii, 1978; Wolfe, Martin, & Palmer, 2000; Wolfe & Steinfatt, 1987; Yiu & NG, 2004; Zraick & Liss, 2000; Zraick, Liss, Dorman, Case, LaPointe, & Beals, 2000). There is evidence that speech-language pathology graduate students differ in voice ratings as compared to speech-language pathologists with experience in voice disorders. In a study conducted by de Bruijn and Whiteside (2007), voice quality ratings of vocal creak, instability, deviation, and hypo/hyperfunctionality were significantly different between speech-language pathologists with two years of experience and speech-language pathology graduate students. Given evidence that two years of experience leads to differences in judgment, speech-language pathology graduate students should not be considered experienced listeners in the area of voice. As a result, one cannot conclude that perceptions of voice are doubtful, but a true difference in experience, or variability in background, affects internal standards.

1.2. Rating Scale

The manner in which listeners are asked to rate the signal can also affect reliability (Kreiman et al., 2007; Shrivastav et al., 2005). There are three different types of ratings scales that are commonly used throughout the voice perception literature: equal appearing interval scales (EAI), visual analog scales (VAS), and direct magnitude estimation (DME). EAI scales
are a discrete point scale covering anywhere from 5-12 points of equal distance in which a listener makes a metathetic measurement, which measures substitutive changes in regard to quality (Eadie & Doyle, 2002; Yiu & NG, 2004; Zraick & Liss, 2000). VAS and DME are continuous scales which are prothetic measurements, which measure additive changes in regard to magnitude (Yiu & NG, 2004; Zraick & Liss, 2000). Using DME, a listener will compare stimuli by entering a number that relates to the previous stimulus (Eadie & Doyle, 2002; Zraick & Liss, 2000). VAS is a continuous scale with a slider to indicate one’s rating of voice quality. This scale allows a listener to utilize the entire scale and does not include number indicators to aid with placement of the slider.

Research shows that test-retest agreement was significantly higher for ratings using continuous scales versus equal appearing interval scales (Kreiman et al., 2007). Also, Zraick and Liss (2000) indicate the benefit of continuous scales, DME and VAS, as these scales do not assume linearity of voice quality perceptions. In fact, when listeners were using both VAS and EAI to rate the same voice qualities, results indicated that judgments made by the listeners using EAI were skewed and both scales were only moderately correlated (Yiu & NG, 2004). The authors concluded that listeners were not using the scales in the same way. Also, it was hypothesized that listeners may only be using a portion of the EAI scale during rating tasks (Eadie & Doyle, 2002). Only one study compared continuous scales, DME and VAS, for ratings of hypernasality. Reliability of severity ratings were highly correlated between both scales and the author concluded both scales were valid (Cheng, 2006).

1.2.1. Multidimensional Scaling

Some researchers feel that continuous scales are superior because equal appearing interval scales force a listener to make unidimensional ratings on a multidimensional signal
(Kreiman et al., 2007). In fact, Bassich and Ludlow (1986) show that a training session did not help listeners obtain an agreement greater than a 0.80 when using a 7-point equal appearing interval scale for a variety of vocal qualities. Listeners are often using multiple dimensions when judging/rating voice qualities, such as: intensity, harmonics-to-noise ratio, fundamental frequency, jitter, shimmer, etc. (Bassich & Ludlow, 1986; Kempster et al., 1991; Kreiman et al., 2004; Murry et al., 1977).

Through the use of multidimensional scaling (MDS), listeners are only asked to rate the similarity between a pair of stimuli, minimizing individual bias (Wolfe et al., 2000). This allows the researcher to explore the dimensions within the acoustical signal used to make judgments (Kreiman et al., 1994; Wolfe et al., 2000). The difficulty in choosing an appropriate rating scale is eliminated as dimensions are determined by the stimuli and not the scale. INDSCAL (Carroll & Chang, 1970), or individual differences scaling, is often used because it identifies the dimensions representing the underlying perceptual structure for judgments made by each participant as well as judgments made by groups of listeners (Murry et al., 1977; Shrivastav, 2006).

Although continuous scales have been found to have better agreement, researchers continue to use EAI scales during the MDS task (Kreiman & Gerratt, 1996; Kreiman et al., 1994; Kreiman et al., 1990; Kreiman et al., 1992; Murry et al., 1977; Shrivastav, 2006; Zraick et al., 2000). In turn, ratings may be skewed or participants may not be utilizing the entire scale during perceptual tasks. In summary, speech-language pathology graduate students, or groups of listeners with mixed levels and types of experience, are often asked to make unidimensional judgments on a multidimensional signal. These results are then correlated to obtained acoustical measures to determine possible relationships among subjective and objective measures of voice
quality. As a result, not only do results vary in regard to perceptual judgments of voice but research regarding objective measures related to these perceptions also remains contradictory.

1.3. Acoustical Measures

Acoustical measurements provide more objective information during evaluation and treatment and confirm perceptions of voice quality (Gerratt et al., 1993), refining initial perceptions (Fitch, 1990; Orlikoff et al., 1999). Orlikoff et al. (1999) suggested that not only does experience affect one’s perceptions of voice quality but also one’s capacity to accurately pair subjective and objective measures of voice.

The most frequently used acoustical measurements for diagnosis and treatment are measurements of perturbation and/or additive noise (Deal & Emmanuel, 1978; Michaelis, Frölich, & Strube, 1998). Period and amplitude perturbation measures occur frequently in the literature in relation to measures of aperiodicity (Hillenbrand, 1987; Martin, Fitch, & Wolfe, 1995; Michaelis et al., 1998). These measures, as well as noise to harmonics ratio (NHR), have moderate to strong correlations with dysphonic voice quality (Awan & Roy, 2006; Deal & Emmanuel, 1978; Eskenazi, Childers, and Hicks, 1990; Hillenbrand, 1988; Leiberman, 1963; Michaelis et al., 1998; Rontal, Rontal, Jacob, and Rolnick, 1983, as cited by Fitch, 1990; Smith et al., 1978; Wendahl, 1966 as cited by Fitch, 1990; Wolfe and Martin, 1997).

Overall, research determining which acoustical measurements correlate with which perceptual judgments of voice quality have been contradictory; however, there is a well documented relationship between perturbation measurements and disordered voice (Bhuta, Patrick, & Garnett, 2004; Deal & Emmanuel, 1978; Eskenzai, et al., 1990; Hecker & Kruel, 1970; Lieberman, 1963; Martin et al., 1995; Smith et al., 1978; Wolfe & Martin, 1997; Wolfe & Steinfatt, 1987; Yumoto, Sasaki, & Okamura, 1984). There is also a well documented
relationship between noise components and disordered voice (Parsa & Jamieson, 2001; Prosek, Montgomery, Walden, & Hawkins, 1987; Yumoto, Gould, & Baer, 1982). There has been little agreement regarding specifics as to which acoustical measurements correlate with which perceptual judgments of voice quality (Bhuta et al., 2004; Deal & Emmanuel, 1978; Eskenzai, et al., 1990; Hecker & Kruel, 1970; Lieberman, 1963; Martin et al., 1995; Parsa & Jamieson, 2001; Prosek et al., 1987; Smith et al., 1978; Wolfe & Martin, 1997; Wolfe & Steinfatt, 1987; Yumoto et al., 1982; Yumoto et al., 1984). With little agreement among authors, the exact relationship between acoustical measures and perceptions of voice quality remains unknown (Michaelis et al., 1998).

The limited agreement regarding a correlation between acoustical measurements and common perceptions of voice quality may be due to an inconsistent use of anchors, an inconsistent use of various acoustical analysis programs, various types of rating scales, and questionable group selection for experienced listeners. Also, most studies did not include a control group. Aside from type and level of experience and type of rating scale, the type of stimuli presented during the rating task may also be affecting perceptions of voice quality and their relationship to objective measures.

1.4. Stimulus Factors

Acoustical measurements, specifically perturbation measurements and noise to harmonics ratio, are highly correlated with one another (Hillenbrand, 1988) giving evidence of the limitations of using natural voices as stimuli. These correlations make a relationship between a single acoustical measure and a single perception of voice quality impossible to find because a rise in any one acoustical measure may be due to the presence of another. Also with the use of natural voices, one cannot separate high frequency noise in the spectrum (breathiness) from low
frequency noise in the spectrum (amplitude and period perturbation) (Kreiman, Gerratt, & Antoñanzas-Barroso, 2006).

1.4.1. **Synthesized Stimuli**

By systematically controlling one acoustical aspect of the signal at a time, through the use of synthesized stimuli, a researcher can examine possible correlations between perceptions of voice quality and objective measurements of voice (Hillenbrand, 1988; Moore, 1989). By using synthesized vowels, a strong correlation between rough voice quality and jitter and shimmer, and a strong correlation between additive noise and breathiness were revealed (Hillenbrand, 1988). Analysis by synthesis using sustained vowel stimuli is an effective way of controlling only one variable in the signal without affecting others.

1.4.2. **Stimulus Length**

The length of stimuli can also affect perception. Sustained vowels are easily controlled and unaffected by coarticulation and prosody because the vocal tract posture and articulatory postures are held constant (de Krom, 1994; Murry & Doherty, 1980). Although an argument can be made that connected speech is more representative of true speech and listener perceptions of voice quality (Parsa & Jamieson, 2001; Qi, Hillman, & Milstein, 1999), Schoentgen (1989) found sustained vowel stimuli and continuous speech stimuli yielded no differences in perturbation measurements. Parsa and Jamieson (2001) suggested that voice onset and offset as well as dynamic changes during connected speech tasks may affect period and amplitude perturbations, respectively. They also compared acoustical measurements from both sustained vowels and connected speech and found that jitter and shimmer measurements obtained from sustained vowel samples were more reliable and better indicators of abnormal voice quality.
1.5. Summary

1.5.1. Interrater Agreement

As discussed, level and type of experience (internal standards), type of rating scale, and type and length of stimuli have been shown to affect listener perceptions. With little control over all possible variables affecting these perceptions, it is no surprise that many researchers concluded that subjective measures of voice quality cannot be trusted. In a study conducted by Kreiman and Gerratt (1996), expert listeners demonstrated good intrarater reliability but poor interrater reliability. In fact, interrater reliability was still poor when expert listeners made categorical judgments for perceptions of breathiness and roughness (Kreiman & Gerratt, 2000). This is of concern as expert listeners are the individuals treating voice disorders and most would assume they have the same academic background. Much of the literature indicates that poor interrater reliability among listeners may be due to individual differences among listeners such as bias and error (Eadie & Doyle, 2002; Kreiman & Gerratt, 1996). Not only were groups not always clearly defined, but there was a lack of a control group, a difference in the use of various ratings scales, and an inconsistency in the use of various types and length of stimuli across studies of voice perception. In conclusion, poor interrater reliability may be secondary to research design versus a true lack of reliability in perceptions of voice (Kreiman et al., 2007).

1.5.2. Standardized Rating Scales

Interrater reliability remains poor even with the use of standardized rating scales, such as GRBAS (Hirano, 1981) and Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) (Kempster, Gerratt, Verdolini Abbott, Barkmeier-Kraemer, & Hillman, 2009). When using GRBAS (Hirano, 1981), which is a categorical scale judging ratings of grade (overall severity), roughness, breathiness, asthenia (weakness), and strain on a scale from 0 (normal) to 3 (severe),
experienced listeners (in this case individuals with mixed levels and types of experience) demonstrated only a moderate test-retest reliability (De Bodt et al., 1997). Also, Sellars, Stanton, McConnachie, Dunnet, Chapman, Bucknell, and Mackenzie (2009) found that experienced listeners only demonstrated an interrater reliability of 0.647, describing this as “rather modest.” Listeners in this study were experienced speech-language pathologists who used GRBAS throughout their practice and were asked to discuss their understanding of the rating scale prior to the experiment. In contrast, when using the CAPE-V (Kempster et al., 2009), a continuous rating scale for judgments of overall severity, roughness, breathiness, pitch, and strain, Kelchner et al. (2010) found experienced speech-language pathologists demonstrated what they considered a strong interrater reliability (0.67 or greater). Both sets of authors, however, describe nearly the same percentage of interrater reliability in completely different ways, leaving the reader unsure how to interpret the results. All of the listeners in both studies were said to be experienced with rating dysphonia, implying a familiarity with the scales being used. In addition, it may be important to note that Kelchner et al. (2010) only used three speech-language pathologists as listeners.

These two standardized ratings scales have been compared and found to only moderately agree (Karnell, Melton, Childes, Coleman, Dailey, & Hoffman, 2007). However, in this study clinicians completed both scales at the same time which could have had an impact on the results. Lastly, Zraick et al. (2011) also had speech-language pathologists with experience with dysphonia rate voice qualities using both GRBAS and CAPE-V. Although reliability varied across both scales, there was a slight improvement in reliability of ratings using the CAPE-V, a continuous scale.
1.5.3. Perceptions of Voice Quality

Despite previous findings, perceptions of voice quality continue to be used and deserve continued attention. Voice quality carries a large amount of information for a communication partner such as: emotional status, personality, communicative intent, geographical background, and gender (Kreiman, Van Lancker-Sidtis, & Gerratt, 2005). An individual experiencing a voice disorder may convey a message he or she did not intend (Kreiman et al., 2005).

Subjective measures of voice quality are what lead many individuals to seek voice treatment (Fastl, 2005; Kreiman et al., 1993). Clinicians/therapists also rely on their ears to guide evaluation and treatment progress. These perceptual judgments, often called the “gold standard,” are used by many voice professionals and help to relate subjective and objective measures of voice (Bassich & Ludlow, 1986; Colton & Casper, 1996; Eadie & Doyle, 2002; Karnell et al., 2007; Kent, 1996; Kreiman & Gerratt, 1998; Kreiman & Gerratt, 2000; Kreiman et al., 1993; Oates, 2009; Orlikoff et al., 1999, p. 4).

1.5.4. Précis

Although there have been many studies in the area of perceptions of voice quality, variables such as the use of anchors, type of rating scale, type and length of stimuli, and level and type of experience can affect perceptual judgments of voice quality, and then in turn, affect correlations with acoustical measurements of voice. Despite our knowledge of these factors affecting perceptions of voice quality, there are very few studies controlling for all of the above variables simultaneously. Most importantly, there are very few studies that address the differences between experienced and inexperienced listeners for perceptions of voice quality. Additionally, generalizations are made for appropriate rating scales and correlations among
acoustical measurements and perceptions through the common use of speech-language pathology graduate students as judges.

One must determine the differences between experienced and inexperienced listeners, if this factor has been found to affect internal standards, before generalizing results. A consistent lack of control for variables affecting perceptions may be the reason for frequent disagreement among authors in the literature. In turn, careful group selection may yield different results for correlations among acoustical components of the signal and perceptions, appropriate rating scales during perceptual tasks, and appropriate length and type of stimuli.

1.6. Purpose

The purpose of this study was to determine the effect of experience on the judgment of voice quality, to determine the perceptual space being used across these groups when judging voice quality, and to examine the correlation among acoustical measurements and perceptions of voice quality.

1.7. Research Questions

1. Is there a significant difference in ratings of perceptions of voice quality between experienced listeners and inexperienced listeners when systemically altering acoustical measurements?

2. Is there a significant difference in ratings of perceptions of voice quality between graduate students and inexperienced listeners when systemically altering acoustical measurements?

3. Is there a significant difference in ratings of perceptions of voice quality between speech-language pathologists and singing voice teachers when systemically altering acoustical measurements?
4. What perceptual changes occur when systematically altering jitter and shimmer, and noise to harmonics ratio (NHR)?

5. What acoustical components of the signal correlate with ratings of voice quality for experienced and inexperienced listeners?

6. Are there differences in the perceptual spaces between experienced and inexperienced listeners?

The null hypothesis was that ratings of voice quality with systematically altered acoustical variables were independent from experience and there was no correlation between any one acoustical variable and the perception of voice quality among any groups. The alternative hypothesis was that ratings of voice quality with systematically altered acoustical variables were dependent upon experience and there was a correlation between acoustical variables and the perception of voice quality among groups.
2. CHAPTER TWO - METHODS

Two experiments were conducted to determine the effect of experience on the perceptions of voice quality and the correlation between acoustical measurements and those perceptions. Experiment One examined the effect of experience when perceiving voice qualities with systematically altered acoustical measurements of jitter, shimmer, and NHR. Experiment Two used multidimensional scaling (MDS) to control for rating scale as listeners were only asked to make judgments about similarities between pairs of stimuli. As a result, the dimensions were determined by the stimuli and not by the scale.

2.1. Experiment One

2.1.1. Stimuli

One sample of sustained vowel /a/ with normal voice quality obtained from a female, age 23, was synthesized using the UCLA synthesizer (Kreiman et al., 2006). This sample, originally recorded at the University of Utah, was chosen because of its widespread use in other studies as an anchor to control for internal standards. Also, the sample was judged to be “normal” by speech-language pathologists who have experience in the area of voice and voice disorders on the basis of quality, pitch and loudness (Awan & Roy, 2005; 2006; 2009a; 2009b). This file, following synthesis, was analyzed using PRAAT (Boersma & Weenink, 2009), a freely available computer acoustical analysis software system, and found to have a mean fundamental frequency of 204.07 Hz, a fundamental frequency standard deviation of 0.04 Hz, a jitter measurement (relative average perturbation percentage) of 0.05, a shimmer in dB measurement of 0.25, and a noise-to-harmonics ratio 0.0036. All obtained measurements were within normal limits for the sex and age of this recorded individual (Baken, 1996).
The UCLA Voice synthesizer (Kreiman et al., 2006) allows one to alter many parameters of voice associated with common perceptual qualities for vowel production. Based on the source-filter theory, this formant synthesizer is limited to sustained vowels and follows a Liljencrants and Fant model (Fant, Liljencrants, & Lin, 1985; Kreiman et al., 2006). The source-filter theory states the source (vocal folds) and filter (vocal tract) are independent of one another (Fant, 1960); therefore, each can be examined and manipulated separately. The acoustic signal is a complex sum of the glottal source, the vocal tract transfer function, and the radiation characteristic (Kreiman et al., 2006). By removing the vocal tract transfer function and the radiation characteristic, one ultimately obtains the glottal flow derivative (Fant, 1979). Linear Predictive Coding, or LPC analysis, is used to estimate the resonant frequencies of the vocal tract and by subtracting this information from the input signal, only the glottal source remains (Kreiman et al., 2006). The glottis is a volume velocity source and volume velocity associates vocal fold movements with the acoustical output (Rothenberg, 1972). The volume velocity waveform can be obtained by inverse-filtering the acoustical waveform (Rothenberg, 1972). An output pulse with little residual, or an absence of remaining formant effects, is considered an appropriate inverse filtered waveform (Kreiman et al., 2006).

The file was downsampled from a 25 kHz to a 10 kHz sampling rate for the inverse filter application. Next, 8 cycles were selected and a spectrum was obtained using Fast Fourier Transform and LPC analysis estimated the resonances of the vocal tract. Then, the fundamental frequency was calculated, followed by inverse filtering, obtaining the glottal flow derivative. The following adjustments can be made to obtain a flow derivative without excessive ripple: adding or removing a resonance, moving a resonance, manipulation of bandwidths, and/or canceling of spectral zeros (Kreiman et al., 2006). However, no adjustments were needed for
this particular file resulting in an appropriate derivative (without excessive ripple) and the file was saved for synthesis.

The LF model (Fant et al., 1985) provides a fast estimate the glottal flow derivative from the inverse filtered acoustical waveform. This model is then fit to the output of the inverse filter; however, this output cannot always accurately model a disordered voice quality and adjustments to the time and frequency domains, such as spectral adjustments and alterations in periodicity, can be made in the synthesizer to portray various voice qualities (Kreiman et al., 2006).

Again, a sampling rate of 10kHz was used to synthesize the inverse filtered file. Samples, with a duration of 1 second, were created with a constant fundamental frequency and amplitude. Then, systematic changes were made to jitter, shimmer, and NHR ratio and synthesized again. Lastly, output from the synthesizer was in ASCII format, so Cool Edit Professional (Cool Edit Pro, 1997) was used to amplify the generated files and convert the files to RIFF format.

For Experiment One, the newly synthesized file was systematically altered by changing measurements of jitter, shimmer, and NHR to create two sets of stimuli. The first set of stimuli included variations of jitter and shimmer simultaneously in evenly spaced intervals (5 intervals) resulting in 25 stimuli. Using the UCLA synthesizer, (Kreiman et al., 2006), jitter covers a range from 0 to 3 µs and shimmer covers a range from 0 to 2 dB. Jitter was altered in increments of 0.75 µs and shimmer was altered in increments of 0.5 dB. The second set of stimuli included a variation of NHR in evenly spaced intervals (10 intervals) resulting in 10 stimuli. Again, using the UCLA synthesizer, NHR covers a range of 0 dB to 50 dB. NHR was altered in increments of 5 dB.
Jitter and shimmer were systematically altered in 5 intervals for a total of 25 combinations and NHR was systematically altered in 10 intervals for a total of 10 stimuli. Combining jitter/shimmer stimuli and NHR stimuli resulted in 35 total stimuli. Jitter, shimmer, and NHR combination stimuli were not generated for this study. Previous studies have shown a strong correlation between noise measurements and perturbation measurements when using natural stimuli (Hillenbrand, 1987; Michaelis et al., 1998). By using synthesized sustained vowels to control only one dimension at a time, Hillenbrand (1988) found a strong nonlinear relationship between period and amplitude perturbations and roughness and a strong correlation between additive noise and breathiness. Therefore, NHR was separated from perturbation measurements for this experiment. Aside from contributing to different types of noise in the spectrum possibly affecting judgments of voice quality, altering jitter and shimmer simultaneously and NHR independently represented the most common and specific perceptual qualities of voice (roughness and breathiness) based on the literature to date (Maryn, Roy, De Bodt, Van Cauwenberge, & Corthals, 2009; Kreiman et al., 2006). Lastly, in an effort to control for fatigue during the experiment, aperiodity and additive noise components were altered separately to significantly reduce the number of stimuli from 250 samples to 35 samples.

2.1.2. Listeners

There were 6 groups with 10 listeners in each group (n=60). Groups consisted of 10 speech-language pathologists, 10 singing voice teachers, 10 speech-language pathology graduate students who had completed a voice disorders course and had not had a voice client, 10 speech-language pathology graduate students who had completed a voice disorders course and had treated one or more voice clients, 10 graduate students in the music department who had completed a voice pedagogy course, and 10 inexperienced listeners.
Group one consisted of 7 females and 3 males who were American Speech Language Hearing Association certified and state licensed speech-language pathologists. Ages ranged from 29-67 years (M=45.7, SD=12.92). They had a range of 5-35 years of experience in voice disorders (M=19, SD=11.01) and spent 10-40 hours per week in the area of voice disorders (M=23.4, SD=12.21). All participants in group one reported no history of a hearing loss, a language disorder, a speech impairment, and/or a neurological disorder.

Group two consisted of 8 females and 2 males, ages ranging from 48-69 years (M=59.6, SD=6) who were tenured singing voice faculty and full members of the National Association of Teachers of Singing (NATS). Individuals holding a full membership in NATS, with either a Master’s Degree or Doctor of Musical Arts (DMA), teach an average of six or more singing voice students weekly with two or more years of experience (NATS, n.d.). The criterion of tenure implies at least six years of full-time faculty work in which the individual mentors undergraduate and graduate students throughout their academic degree of study. Mentoring may include (but is not limited to) guidance in the following areas: providing weekly singing lessons, student auditions, student performances, voice jury examinations, student competitions, and student degree recitals. The criterion for obtaining tenured singing voice faculty, rather than just a degree specification, was established because the Doctorate of Musical Arts degree may not always imply the above listed experiences. Typically, in degree work, the student will focus most of his or her time on their own singing versus teaching and analyzing the voices of others. All participants in group two reported no history of a hearing loss, a language disorder, a speech impairment, and/or a neurological disorder.

These definitions of an experienced listener, for both speech-language pathologists and singing voice teachers, have been shown to result in different levels of agreement when
classifying voice quality (Sofranko & Prosek, 2012). In the current experiment, the inclusion criteria for a singing voice teacher was more clearly defined than Sofranko and Prosek (2012), by including tenured voice faculty. This was to indicate a specific type of experience in an effort to control for additional variability in the academic background and experience of the singing voice teachers.

Group three consisted of 10 females, ages ranging from 21-24 years (M=22, SD=0.943), who were current graduate students in a speech-language pathology program and had completed a voice disorders course. Group four, although similar, consisted of 10 females, ages ranging from 21-42 years (M=26.1, SD=6.33), who were also current graduate students in a speech-language pathology program, had completed a voice disorders course, but these students had also had one or more voice client(s) in clinic. Students had a range of 1-8 voice clients in their clinical experience (M=2.5, SD=2.321). These two groups were included to examine any possible differences between students who had a clinical experience in the area of voice as compared to those who had not, following the same academic coursework. All participants in groups three and four reported no history of a hearing loss, a language disorder, a speech impairment, and/or a neurological disorder.

Group five consisted of 6 females and 4 males with ages ranging from 22-46 years (M=27.9, SD=7.4). These individuals were current graduate students in either voice pedagogy or vocal performance who had completed a voice pedagogy course in their graduate work. The students taught a range of 1-20 singing voice students weekly (M=5.6, SD=5.48). All participants in group five reported no history of a hearing loss, a language disorder, a speech impairment, and/or a neurological disorder. This group was included to parallel the graduate speech-language pathology student groups to examine the effect of level of experience on
judgments of voice quality when comparing the experienced listener groups included in the study.

Lastly, group six consisted of 5 females and 5 males, ages ranging from 24-56 years (M=35, SD=12.18), with no previous training in voice and/or voice disorders, including singing lessons and voice treatment. This group included individuals from various backgrounds including: nursing, real estate, chemistry, culinary arts, fashion, architecture, cosmetology, engineering (mechanical and electrical), and law. All participants in group six reported no history of a hearing loss, a language disorder, a speech impairment, and/or a neurological disorder.

2.1.3. Procedures

Detailed instructions were provided prior to beginning the experiment defining voice qualities of overall severity, roughness, breathiness, pitch adequacy, and strain (See Appendix A). Overall severity was defined as “a global, integrated impression of voice deviance” (CAPE-V, Kempster et al., 2009, p.130); breathiness was defined as the escape of air during voicing that one can hear and which reduces loudness (Colton & Casper, 1996); roughness was defined as a voice quality lacking clarity with increased noisiness characterized by tension or strain one can hear (Colton & Casper, 1996); pitch adequacy was defined as a rating to determine if a “pitch deviates from normal for that person’s gender, age, and referent culture” (CAPE-V, Kempster, et al., 2009, p.130) and strain was defined as difficulty maintaining and controlling voicing with moments of voice loss (Colton & Casper, 1996). These five voice qualities were selected for this study because they are present on both the GRBAS (Hirano, 1981) and CAPE-V (Kempster, et al., 2009) standardized rating scales. However, these qualities more closely parallel the CAPE-V standardized scale because it is also a continuous rating scale.
Participants listened to the synthesized samples and rated each voice quality on a visual analog scale ranging from mild to severe with three indicators “MI,” for mild, “MO,” for moderate, and “SE,” for severe, included on the scale for supplemental severity indicators (Kempster et al., 2009), covering a range from 1-1000, for the following categories: overall severity, breathiness, roughness, strain, and pitch adequacy. The range of numbers was related to data collection for the visual analog scale, but was not indicated on the scale during the experiment (See Figure 2.1). The stimuli were presented and the results were collected using Alvin2 (Hillenbrand, 2005). Playback level was adjusted to a comfortable level for each participant individually.

Figure 2.1. Visual analog screen display in Alvin2.

Each stimulus was presented twice to evaluate intrarater reliability (n=70). Stimuli were presented in random order for each participant to control for order effects. No feedback was provided and listeners were able to replay stimuli as needed throughout the study.
2.2.** Experiment Two**

2.2.1. **Stimuli**

The same synthesized jitter/shimmer combinations (n=25) used for experiment one were used in experiment two; however, systematically altered NHR stimuli were changed from 10 stimuli to 5 stimuli including intervals of 12.5 dB rather than 5 dB. This allowed for 435 pairs of stimuli to be presented, rather than 595 pairs of stimuli in an effort to better prevent fatigue. Experiment two examined the perceptual spaces of the listeners during a multidimensional scaling task.

2.2.2. **Listeners**

The same 60 listeners who completed Experiment One also completed Experiment Two. Five participants in each of the five groups completed Experiment One first while the other five participants completed Experiment Two first. In either case, there was a 10 minute break in between experiments in which participants filled out a survey targeting personal recreational information and/or voice background information to eliminate a possible learned orientation to the stimuli and allow for reorientation to a different task.

2.2.3. **Procedures**

Experiment two used multidimensional scaling (MDS), specifically INDSCAL (Carroll & Chang, 1970) to examine dimensions within the signal used to make those judgments. INDSCAL, or individual differences scaling, has been used to examine perceptions of voice quality because it identifies dimensions that represent the underlying structure of judgments of groups of listeners without averaging data across participants (Murry et al., 1977; Shrivastav, 2006). Listeners were randomly presented with pairs (n=435) of the generated stimuli auditorily. All intraorder pairs were not included (i.e. AB and BA) as research has shown that order does
not influence the judgments of listeners (Mohr & Wang, 1968; Walden, Montgomery, Gibeily, Prosek, & Schwartz, 1978). Detailed instructions were provided prior to the experiment (See Appendix B). Participants were asked to rate the similarity between voice samples on a visual analog scale, ranging from no similarity to extremely similar, covering a range from 1 to 1000. Again, the range of numbers was related to data collection for the visual analog scale, but was not indicated on the scale during the experiment (See Figure 2.2). Participants were not permitted to replay samples during this task and there was no orientation task or training session prior to the experiment, following an experimental procedure for multidimensional scaling in the area of voice perception (Kreiman & Gerratt, 2005; Kreiman et al., 1994; Kreiman et al., 1990). The stimuli were presented and the results were collected using Alvin2 (Hillenbrand, 2005). Playback level was adjusted to a comfortable level for each participant individually.

*Figure 2.2. Multidimensional scaling screen layout in Alvin2.*
3. CHAPTER THREE – RESULTS

3.1. Experiment One

3.1.1. Intrarater Agreement

Minitab® Statistical Software was used to calculate intrarater agreement for ratings of overall severity, roughness, breathiness, strain, and pitch using t-tests comparing the first and second ratings of each stimulus across all groups (Minitab, 2005). P-values greater than an alpha of 0.01, following a Bonferroni correction, indicate that the means for the listener’s first rating did not significantly differ from the means for the listener’s second rating. Overall, speech-language pathologists (SLPs), graduate speech-language pathology graduate students (who have and have not had a voice client)(SLPGRADVs and SLPGRADs respectively), graduate students who have completed a voice pedagogy course (SVTGRADs), and inexperienced listeners (IEs) did not differ significantly from rating one to rating two across all scales. Tenured singing voice faculty (SVTs) performed similarly from rating one to rating two for breathiness, strain, and pitch, but did differ significantly from rating one to rating two for overall severity, and roughness. (See Table 3.1)
### Table 3.1 Summary of Intrarater Agreement

<table>
<thead>
<tr>
<th>Group</th>
<th>Overall Severity</th>
<th>Roughness</th>
<th>Breathiness</th>
<th>Strain</th>
<th>Pitch</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLPs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t-stat</td>
<td>$t(697) = -1.28$</td>
<td>$t(697) = 0.46$</td>
<td>$t(697) = -1.06$</td>
<td>$t(697) = -0.54$</td>
<td>$t(696) = 1.18$</td>
</tr>
<tr>
<td>p-value</td>
<td>$p&gt;0.01$</td>
<td>$p&gt;0.01$</td>
<td>$p&gt;0.01$</td>
<td>$p&gt;0.01$</td>
<td>$p&gt;0.01$</td>
</tr>
<tr>
<td>SLPGRADs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t-stat</td>
<td>$t(697) = 0.51$</td>
<td>$t(697) = 1.23$</td>
<td>$t(697) = 0.64$</td>
<td>$t(697) = 1.84$</td>
<td>$t(693) = 0.64$</td>
</tr>
<tr>
<td>p-value</td>
<td>$p&gt;0.01$</td>
<td>$p&gt;0.01$</td>
<td>$p&gt;0.01$</td>
<td>$p&gt;0.01$</td>
<td>$p&gt;0.01$</td>
</tr>
<tr>
<td>SLPGRADVs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t-stat</td>
<td>$t(697) = -0.31$</td>
<td>$t(697) = 0.23$</td>
<td>$t(697) = 0.53$</td>
<td>$t(697) = -0.97$</td>
<td>$t(690) = 1.18$</td>
</tr>
<tr>
<td>p-value</td>
<td>$p&gt;0.01$</td>
<td>$p&gt;0.01$</td>
<td>$p&gt;0.01$</td>
<td>$p&gt;0.01$</td>
<td>$p&gt;0.01$</td>
</tr>
<tr>
<td>SVTs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t-stat</td>
<td>$t(697) = 2.97$</td>
<td>$t(697) = 2.60$</td>
<td>$t(697) = 1.08$</td>
<td>$t(697) = 1.51$</td>
<td>$t(696) = 2.07$</td>
</tr>
<tr>
<td>p-value</td>
<td>$p&lt;0.01$*</td>
<td>$p&lt;0.01$*</td>
<td>$p&gt;0.01$</td>
<td>$p&gt;0.01$</td>
<td>$p&gt;0.01$</td>
</tr>
<tr>
<td>SVTGRADs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t-stat</td>
<td>$t(697) = 0.40$</td>
<td>$t(697) = 0.55$</td>
<td>$t(697) = -0.21$</td>
<td>$t(697) = -0.01$</td>
<td>$t(696) = -0.57$</td>
</tr>
<tr>
<td>p-value</td>
<td>$p&gt;0.01$</td>
<td>$p&gt;0.01$</td>
<td>$p&gt;0.01$</td>
<td>$p&gt;0.01$</td>
<td>$p&gt;0.01$</td>
</tr>
<tr>
<td>IEs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t-stat</td>
<td>$t(697) = 1.16$</td>
<td>$t(697) = 1.33$</td>
<td>$t(697) = 0.82$</td>
<td>$t(697) = 1.55$</td>
<td>$t(696) = -0.58$</td>
</tr>
<tr>
<td>p-value</td>
<td>$p&gt;0.01$</td>
<td>$p&gt;0.01$</td>
<td>$p&gt;0.01$</td>
<td>$p&gt;0.01$</td>
<td>$p&gt;0.01$</td>
</tr>
</tbody>
</table>

Note. Speech-language pathologists (SLPs), speech-language pathology graduate students who have not had a voice client (SLPGRADs), speech-language pathology graduate students who have had a voice client (SLPGRADVs), tenured singing voice faculty (SVTs), graduate students who have completed a voice pedagogy course (SVTGRADs), and inexperienced listeners (IEs) and * indicates lack of agreement between rating one and rating two.

#### 3.1.2. Intrarater Agreement

Minitab® Statistical Software was used to calculate intrarater agreement for ratings of overall severity, roughness, breathiness, strain, and pitch using averaged Pearson $r$ correlations (Minitab, 2005). Correlations were tested against zero and p-values less than an alpha of 0.01, following a Bonferroni correction, and indicated that listeners in that group were agreeing with one another across judgments of overall severity, roughness, breathiness, strain, and pitch.

None of the obtained correlations accounted for at least 50% of the variance, but all correlations for overall severity, roughness, and breathiness were significant for expert listeners and student groups. When put in order from highest agreement to lowest agreement for ratings of overall severity and roughness, SLPs followed by the SVTs had the highest levels of intrarater agreement. Student groups and inexperienced listeners followed with lesser levels of intrarater agreement.
agreement. The correlation for inexperienced listeners was insignificant for ratings of overall severity. For breathiness, SLP graduate student groups had the highest levels of interrater agreement followed by SLPs, SVTs and SVTGRADs and IEs with the lowest levels. Both strain and pitch ratings had insignificant, low, interrater agreement levels across all groups. (See Table 3.2)

Table 3.2 Summary of Interrater Agreement

<table>
<thead>
<tr>
<th>Group</th>
<th>Overall</th>
<th>Severity</th>
<th>Roughness</th>
<th>Breathiness</th>
<th>Strain</th>
<th>Pitch</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLPs</td>
<td>Pearson</td>
<td>r = 0.641</td>
<td>r = 0.647</td>
<td>r = 0.535</td>
<td>r = 0.254</td>
<td>r = 0.216</td>
<td>r = 0.459</td>
</tr>
<tr>
<td></td>
<td>p-value</td>
<td>p&lt;0.01</td>
<td>p&lt;0.01</td>
<td>p&lt;0.01</td>
<td>p&gt;0.01*</td>
<td>p&gt;0.01*</td>
<td>p&lt;0.01</td>
</tr>
<tr>
<td>SLPGRADs</td>
<td>Pearson</td>
<td>r = 0.392</td>
<td>r = 0.398</td>
<td>r = 0.571</td>
<td>r = 0.288</td>
<td>r = 0.136</td>
<td>r = 0.357</td>
</tr>
<tr>
<td></td>
<td>p-value</td>
<td>p&lt;0.01</td>
<td>p&lt;0.01</td>
<td>p&lt;0.01</td>
<td>p&gt;0.01*</td>
<td>p&gt;0.01*</td>
<td>p&lt;0.01</td>
</tr>
<tr>
<td>SLPGRADVs</td>
<td>Pearson</td>
<td>r = 0.564</td>
<td>r = 0.588</td>
<td>r = 0.664</td>
<td>r = 0.259</td>
<td>r = 0.207</td>
<td>r = 0.456</td>
</tr>
<tr>
<td></td>
<td>p-value</td>
<td>p&lt;0.01</td>
<td>p&lt;0.01</td>
<td>p&lt;0.01</td>
<td>p&gt;0.01*</td>
<td>p&gt;0.01*</td>
<td>p&lt;0.01</td>
</tr>
<tr>
<td>SVTs</td>
<td>Pearson</td>
<td>r = 0.602</td>
<td>r = 0.631</td>
<td>r = 0.531</td>
<td>r = 0.336</td>
<td>r = 0.262</td>
<td>r = 0.472</td>
</tr>
<tr>
<td></td>
<td>p-value</td>
<td>p&lt;0.01</td>
<td>p&lt;0.01</td>
<td>p&lt;0.01</td>
<td>p&gt;0.01*</td>
<td>p&gt;0.01*</td>
<td>p&lt;0.01</td>
</tr>
<tr>
<td>SVTGRADs</td>
<td>Pearson</td>
<td>r = 0.520</td>
<td>r = 0.526</td>
<td>r = 0.468</td>
<td>r = 0.323</td>
<td>r = 0.316</td>
<td>r = 0.431</td>
</tr>
<tr>
<td></td>
<td>p-value</td>
<td>p&lt;0.01</td>
<td>p&lt;0.01</td>
<td>p&lt;0.01</td>
<td>p&gt;0.01*</td>
<td>p&gt;0.01*</td>
<td>p&lt;0.01</td>
</tr>
<tr>
<td>IEs</td>
<td>Pearson</td>
<td>r = 0.274</td>
<td>r = 0.463</td>
<td>r = 0.408</td>
<td>r = 0.323</td>
<td>r = 0.172</td>
<td>r = 0.323</td>
</tr>
<tr>
<td></td>
<td>p-value</td>
<td>p&gt;0.01*</td>
<td>p&lt;0.01</td>
<td>p&lt;0.01</td>
<td>p&gt;0.01*</td>
<td>p&gt;0.01*</td>
<td>p&lt;0.01</td>
</tr>
</tbody>
</table>

Note. Speech-language pathologists (SLPs), speech-language pathology graduate students who have not had a voice client (SLPGRADs), speech-language pathology graduate students who have had a voice client (SLPGRADVs), tenured singing voice faculty (SVTs), graduate students who have completed a voice pedagogy course (SVTGRADs), and inexperienced listeners (IEs) and * indicates an insignificant correlation.

3.1.3. Correlation

Using Minitab® Statistical Software, Pearson r correlations between types of stimuli and scale rating were calculated for all groups and tested against zero at an alpha level of 0.01 (Minitab, 2005). Although none of the calculated Pearson r correlations accounted for more than 50% of the variance, graphing obtained significant correlations across groups showed some
distinct patterns. Only correlations that were significant with rating scale at an alpha level of 0.01 were plotted.

For ratings of overall severity, jitter/shimmer combination stimuli and shimmer only stimuli consistently correlated to ratings more than jitter only stimuli. Also, SVTs and SVTGRADs demonstrated higher correlations for all three types of stimuli as compared to all other groups. (See Figure 3.1)

**Figure 3.1 Correlations between Types of Stimuli and Overall Severity**

[Figure showing correlations between types of stimuli and overall severity]

*Note.* All Pearson r correlations graphed above were significant at an alpha level of 0.01.

Jitter/shimmer combination stimuli and shimmer only stimuli followed the same pattern across groups for ratings of roughness, having the highest correlations as compared to jitter only stimuli. SVTs had the highest correlation among jitter/shimmer combination stimuli and shimmer only stimuli with ratings of roughness. (See Figure 3.2)
Groups did not demonstrate significant correlations between jitter/shimmer, shimmer only, and jitter only stimuli with ratings of breathiness. Only NHR stimuli had consistent significant correlations with these ratings. Both SLP graduate students groups demonstrated the highest correlation between NHR stimuli and ratings of breathiness. (See Figure 3.3)
Almost every group, except SLPs, had the highest correlations between jitter/shimmer combination stimuli and ratings of strain. Also, aside from SVTGRADs, each group demonstrated that shimmer only stimuli had the second highest correlation when rating strain. (See Figure 3.4)
Figure 3.4 Correlations between Types of Stimuli and Strain

Note. All Pearson r correlations graphed above were significant at an alpha level of 0.01.

Acoustical measures only had significant correlations for ratings of pitch for SVTs, SVTGRADs, and IEs. Across all three groups, jitter/shimmer combination stimuli had the highest correlation. (See Figure 3.5)
3.1.4. Scale Analysis

Minitab® Statistical Software was used to perform a MANOVA with overall severity, roughness, breathiness, strain, and pitch ratings as the response and group (all experienced and inexperienced listeners) and type of stimulus (jitter/shimmer combinations and NHR systematically altered samples) as predictors at an alpha level of 0.05 (Minitab, 2005). Results showed that group, $F(25, 7698) = 14.82, p < 0.05$, stimulus, $F(15, 5720) = 68.24, p < 0.05$, and an interaction between group and stimulus, $F(75, 9928) = 1.79, p < 0.05$, were significant factors. Results were then analyzed, again with Minitab® Statistical Software, using five analyses of variance with overall severity, roughness, breathiness, strain, and ratings as the response and group and type of stimulus as the predictors (Minitab, 2005). An interaction term for group and type of stimulus was also included. Following a Bonferroni adjustment, the alpha
level was 0.01. A Tukey Honestly Significance Difference (HSD) was conducted, as the overall F score was significant for both group and type of stimulus across all five scales.

### 3.1.4.1. Overall severity

Analysis of variance revealed a significant effect for group and type of stimulus on judgments of overall severity, $F(5, 2076) = 14.59, p < 0.01$ and $F(3, 2076) = 41.59, p < 0.01$, respectively. An interaction effect of group and type of stimulus was not significant, $F(15, 2076) = 1.22, p > 0.01$. A Tukey HSD test was conducted to determine difference in means across the groups of listeners and types of stimuli. Tukey’s HSD criterion indicated that SVTs and SVTGRADs performed similarly and were significantly different when rating overall severity than IEs, SLPs, and both SLP graduate student groups. Overall, SVTs and SVTGRADs had the highest ratings of overall severity. (See Table 3.3 and Figure 3.6)

**Table 3.3 Group: Tukey HSD Results for Overall Severity**

<table>
<thead>
<tr>
<th>Group</th>
<th>Mean</th>
<th>SD</th>
<th>Letter</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVTs</td>
<td>519.4</td>
<td>283.8</td>
<td>A</td>
</tr>
<tr>
<td>SVTGRADs</td>
<td>502.2</td>
<td>283.4</td>
<td>A</td>
</tr>
<tr>
<td>IEs</td>
<td>398.2</td>
<td>260.7</td>
<td>B</td>
</tr>
<tr>
<td>SLPs</td>
<td>385.8</td>
<td>267.1</td>
<td>B</td>
</tr>
<tr>
<td>SLPGRADs</td>
<td>381.5</td>
<td>244.2</td>
<td>B</td>
</tr>
<tr>
<td>SLPGRADVs</td>
<td>364.7</td>
<td>269.1</td>
<td>B</td>
</tr>
</tbody>
</table>

*Note.* Means that do not share a letter are significantly different (Minitab, 2005).
A Tukey HSD test was also performed for type of stimulus. Stimuli with altered combinations of jitter and shimmer had the highest mean in relation to ratings of overall severity. Stimuli with altered noise-to-harmonics ratios had the lowest mean in relation to ratings of overall severity. There were significant differences between the highest and lowest means with some relationship among groups for the remaining types of stimuli. Jitter/shimmer combination stimuli were significantly different from jitter only and NHR stimuli and shimmer stimuli were also significantly different from NHR stimuli. (See Table 3.4)

### Table 3.4 Type of Stimulus: Tukey HSD Results for Overall Severity

<table>
<thead>
<tr>
<th>Type of Stimulus</th>
<th>Mean</th>
<th>SD</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Jitter/Shimmer</td>
<td>483.95</td>
<td>253.66</td>
<td>A</td>
</tr>
<tr>
<td>Shimmer</td>
<td>440.30</td>
<td>270.30</td>
<td>A B</td>
</tr>
<tr>
<td>Jitter</td>
<td>369.40</td>
<td>241.20</td>
<td>B C</td>
</tr>
<tr>
<td>NHR</td>
<td>342.00</td>
<td>298.50</td>
<td>C</td>
</tr>
</tbody>
</table>

*Note.* Means that do not share a letter are significantly different (Minitab, 2005).
3.1.4.2. **Roughness**

There was a significant effect for group and type of stimulus on judgments of roughness, $F(5, 2076) = 17.50, p < 0.01$ and $F(3, 2076) = 191.43, p < 0.01$, respectively. An interaction between group and type of stimulus was not significant $F(15, 2076) = 1.94, p > 0.01$. Tukey’s HSD indicated that SVTs, SVTGRADs, and IEs performed similarly in regard to ratings of roughness. These groups had the highest ratings of roughness, with SVTs having the highest rating overall. SLPs and both SLP graduate student groups were significantly different from the SVTs, SVTGRADs, and IEs, with lower ratings of roughness. SLPs had the lowest ratings of roughness overall. (See Table 3.5 and Figure 3.7)

*Table 3.5 Group: Tukey HSD Results for Roughness*

<table>
<thead>
<tr>
<th>Group</th>
<th>Mean</th>
<th>SD</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>SVTs</td>
<td>481.7</td>
<td>327.0</td>
<td>A</td>
</tr>
<tr>
<td>SVTGRADs</td>
<td>432.3</td>
<td>327.1</td>
<td>A</td>
</tr>
<tr>
<td>IEs</td>
<td>417.3</td>
<td>296.0</td>
<td>A</td>
</tr>
<tr>
<td>SLPGRADs</td>
<td>328.2</td>
<td>236.6</td>
<td>B</td>
</tr>
<tr>
<td>SLPGRADVs</td>
<td>332.4</td>
<td>283.1</td>
<td>B</td>
</tr>
<tr>
<td>SLPs</td>
<td>304.6</td>
<td>278.9</td>
<td>B</td>
</tr>
</tbody>
</table>

*Note.* Means that do not share a letter are significantly different (Minitab, 2005).
Again, using a Tukey HSD test, jitter/shimmer combination stimuli and stimuli with altered shimmer were grouped together with the highest mean for ratings of roughness and were significantly different from all other types of stimuli. Stimuli with altered jitter and stimuli with altered NHR were significantly different from one another as well as all other types of stimuli. Lastly, NHR stimuli had the lowest mean in relation to ratings of roughness. (See Table 3.6)

Table 3.6 Type of Stimulus: Tukey HSD Results for Roughness

<table>
<thead>
<tr>
<th>Type of Stimulus</th>
<th>Mean</th>
<th>SD</th>
<th>Letter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jitter/Shimmer</td>
<td>493.58</td>
<td>279.53</td>
<td>A</td>
</tr>
<tr>
<td>Shimmer</td>
<td>444.90</td>
<td>296.00</td>
<td>A</td>
</tr>
<tr>
<td>Jitter</td>
<td>359.70</td>
<td>261.30</td>
<td>B</td>
</tr>
<tr>
<td>NHR</td>
<td>178.72</td>
<td>235.47</td>
<td>C</td>
</tr>
</tbody>
</table>

Note. Means that do not share a letter are significantly different (Minitab, 2005).
3.1.4.3. Breathiness

Analysis of variance revealed another significant effect for group and type of stimulus on judgments of breathiness, $F(5, 2076) = 11.30, p < 0.01$ and $F(3, 2076) = 42.53, p < 0.01$, respectively. An interaction between group and type of stimulus was not significant, $F(15, 2076) = 0.72, p > 0.01$. Results from the Tukey HSD test showed that SVTs had the highest ratings of breathiness and were not significantly different from the SVTGRADs or the IEs. However, SVTs were significantly different than SLPs and both SLP graduate student groups. Individuals with a background in speech-language pathology had the lowest ratings of breathiness overall. In summary, groups who had ratings of breathiness in the mid-range and inexperienced listeners did not differ significantly from any of the other groups. (See Table 3.7 and Figure 3.8)

Table 3.7 Group: Tukey HSD Results for Breathiness

<table>
<thead>
<tr>
<th>Group</th>
<th>Mean</th>
<th>SD</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>SVTs</td>
<td>342.6</td>
<td>319.1</td>
<td>A</td>
</tr>
<tr>
<td>SVTGRADs</td>
<td>319.7</td>
<td>317.2</td>
<td>A B</td>
</tr>
<tr>
<td>IEs</td>
<td>272.9</td>
<td>278.6</td>
<td>A B C</td>
</tr>
<tr>
<td>SLPs</td>
<td>247.6</td>
<td>292.8</td>
<td>B C</td>
</tr>
<tr>
<td>SLPGRADVs</td>
<td>225.0</td>
<td>270.0</td>
<td>C</td>
</tr>
<tr>
<td>SLPGRADs</td>
<td>186.0</td>
<td>235.0</td>
<td>C</td>
</tr>
</tbody>
</table>

*Note.* Means that do not share a letter are significantly different (Minitab, 2005).
Following another Tukey HSD test, stimuli with altered NHR had the highest mean in regard to ratings of breathiness. This mean was also significantly different from all other stimulus types. (See Table 3.8)

Table 3.8 Type of Stimulus: Tukey HSD Results for Breathiness

<table>
<thead>
<tr>
<th>Type of Stimulus</th>
<th>Mean</th>
<th>SD</th>
<th>Letter</th>
</tr>
</thead>
<tbody>
<tr>
<td>NHR</td>
<td>373.4</td>
<td>362.7</td>
<td>A</td>
</tr>
<tr>
<td>Shimmer</td>
<td>227.9</td>
<td>242.1</td>
<td>B</td>
</tr>
<tr>
<td>Jitter/Shimmer</td>
<td>227.3</td>
<td>249.7</td>
<td>B</td>
</tr>
<tr>
<td>Jitter</td>
<td>197.0</td>
<td>225.0</td>
<td>B</td>
</tr>
</tbody>
</table>

Note. Means that do not share a letter are significantly different (Minitab, 2005).

3.1.4.4. Strain

There was a significant effect for group and type of stimulus on judgments of strain, $F(5, 2076) = 27.90, p < 0.01$ and $F(3, 2076) = 53.97, p < 0.01$, respectively. Again, an interaction between group and type of stimulus was not significant, $F(15, 2076) = 1.41, p > 0.01$. 
The Tukey HSD criterion showed, again, that SVTs had the highest ratings and did not differ from the SVTGRADs or IEs. Both the SVTs and SVTGRADs groups differed from SLPs and SLP student groups. The mean obtained for IEs did not differ significantly from SVTGRADs or SLPs and SLPs only differed significantly from their corresponding student group of individuals who had a voice client. Lastly, SLP student groups had the lowest overall ratings of strain. (See Table 3.9 and Figure 3.9)

Table 3.9 Group: Tukey HSD Results for Strain

<table>
<thead>
<tr>
<th>Group</th>
<th>Mean</th>
<th>SD</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>SVTs</td>
<td>420.5</td>
<td>298.3</td>
<td>A</td>
</tr>
<tr>
<td>SVTGRADs</td>
<td>351.8</td>
<td>293.6</td>
<td>A</td>
</tr>
<tr>
<td>IEs</td>
<td>329.6</td>
<td>281.4</td>
<td>A B</td>
</tr>
<tr>
<td>SLPs</td>
<td>256.3</td>
<td>308.0</td>
<td>B C</td>
</tr>
<tr>
<td>SLPGRADs</td>
<td>238.2</td>
<td>200.2</td>
<td>C D</td>
</tr>
<tr>
<td>SLPGRADVs</td>
<td>166.8</td>
<td>209.6</td>
<td>D</td>
</tr>
</tbody>
</table>

*Note. Means that do not share a letter are significantly different (Minitab, 2005).*

Figure 3.9 Mean Ratings for Strain
Stimuli with altered jitter and shimmer combinations had the highest ratings of strain and differed significantly from NHR stimuli, which had the lowest ratings of strain. Stimuli with altered jitter differed significantly from NHR ratio and jitter/shimmer combinations and NHR stimuli different from all other stimulus types. (See Table 3.10)

Table 3.10 Type of Stimulus: Tukey HSD Results for Strain

<table>
<thead>
<tr>
<th>Type of Stimulus</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jitter/Shimmer</td>
<td>354.61</td>
<td>283.77</td>
</tr>
<tr>
<td>Shimmer</td>
<td>313.70</td>
<td>282.70</td>
</tr>
<tr>
<td>Jitter</td>
<td>285.30</td>
<td>267.80</td>
</tr>
<tr>
<td>NHR</td>
<td>186.10</td>
<td>247.00</td>
</tr>
</tbody>
</table>

Note. Means that do not share a letter are significantly different (Minitab, 2005).

3.1.4.5. Pitch

Lastly, an analysis of variance showed there was a significant effect for group and type of stimulus on judgments of pitch, $F(5, 2076) = 35.94, p < 0.01$ and $F(3, 2076) = 23.31, p < 0.01$, respectively. This was the only scale where an interaction between group and type of stimulus was found, $F(15, 2076) = 3.84, p > 0.01$. The Tukey HSD test showed that IEs had the highest ratings for pitch and SLPGRADVs had the lowest ratings for pitch. In fact, SLPGRADVs were the only group that was significantly different from all other groups. Again, SLPs did not differ significantly from SLP graduates who had not had a voice client. Overall, there was some overlap among all of the groups in the mid-range for ratings of pitch. SLPGRADs did not differ significantly from SVTGRADs, who in turn, did not differ significantly from SVTs. (See Table 3.11 and Figure 3.10)
Table 3.11 Group: Tukey HSD Results for Pitch

<table>
<thead>
<tr>
<th>Group</th>
<th>Mean</th>
<th>SD</th>
<th>Letter</th>
</tr>
</thead>
<tbody>
<tr>
<td>IEs</td>
<td>383.70</td>
<td>271.90</td>
<td>A</td>
</tr>
<tr>
<td>SVTs</td>
<td>331.10</td>
<td>274.50</td>
<td>A</td>
</tr>
<tr>
<td>SVTGRADs</td>
<td>293.40</td>
<td>286.50</td>
<td>B</td>
</tr>
<tr>
<td>SLPGRADs</td>
<td>233.10</td>
<td>244.70</td>
<td>C</td>
</tr>
<tr>
<td>SLPs</td>
<td>199.40</td>
<td>239.30</td>
<td>D</td>
</tr>
<tr>
<td>SLPGRADVs</td>
<td>114.22</td>
<td>150.75</td>
<td>E</td>
</tr>
</tbody>
</table>

Note. Means that do not share a letter are significantly different (Minitab, 2005).

Figure 3.10 Mean Ratings for Pitch

Like strain, stimuli with altered jitter and shimmer combinations had the highest ratings and NHR stimuli had the lowest ratings. However, for pitch, NHR stimuli differed from jitter/shimmer combination and jitter stimuli, but not shimmer stimuli. (See Table 3.12)

Table 3.12 Type of Stimulus: Tukey HSD Results for Pitch

<table>
<thead>
<tr>
<th>Type of Stimulus</th>
<th>Mean</th>
<th>SD</th>
<th>Letter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jitter/Shimmer</td>
<td>299.36</td>
<td>261.75</td>
<td>A</td>
</tr>
<tr>
<td>Jitter</td>
<td>269.80</td>
<td>277.40</td>
<td>A</td>
</tr>
<tr>
<td>Shimmer</td>
<td>230.00</td>
<td>245.60</td>
<td>B</td>
</tr>
<tr>
<td>NHR</td>
<td>198.10</td>
<td>256.30</td>
<td>C</td>
</tr>
</tbody>
</table>

Note. Means that do not share a letter are significantly different (Minitab, 2005).
3.1.5. Effect Size

It is customary to calculate effect size to gauge the meaningfulness of the results. The effect size quantifies the size of the difference between means by dividing this difference by the pooled standard deviation. In this way the magnitude of the difference is emphasized rather than the statistical significance. Although the results of this study have significant p-values, the large degrees of freedom reported may lead one to question the magnitude of the relationship among the variables. Partial eta-squared and eta-squared were calculated for the MANOVA and ANOVA results, respectively, for Experiment One. (See Table 3.13 and Table 3.14) All effect sizes were significant at an alpha level of 0.01. According to Cohen (1988), an effect size of 0.01 is considered small, an effect size of 0.06 is considered medium, and an effect size of 0.14 is considered large. The majority of the effect sizes were small, implying both the magnitude of the relationship among judgments of voice quality and stimuli and the magnitude of the relationship among judgments of voice quality and experience were minimal. However, the calculation for degrees of freedom must be reviewed.

Cohen’s (1988) original concern was that a large number of participants in and of itself would lead to statistical significance even when the difference between the means was meaningless. For both MANOVA and ANOVA, the F statistic included numerator and denominator degrees of freedom, \( k - 1 \) and \( n - k \) respectively (\( k \) representing the number of treatments or samples and \( n \) representing the total sample size). When determining \( n \), for this experiment, the calculation included the number of stimuli as well as the number of participants; therefore, effect size was misleading. For Experiment One, there were 60 participants and 35 stimuli. These numbers were multiplied to calculate \( n \) which resulted in large degrees of freedom. Ten participants in each group was an adequate, but not inordinately large, sample
size. One may suggest that degrees of freedom could have been made smaller by decreasing the number of stimuli. However, the number of stimuli cannot be altered. Decreasing the number of stimuli would introduce large gaps in the sampling space, possibly affecting the judgments of the listeners. As a result, the degrees of freedom remained large and effect size was not a reliable gauge of the meaningfulness of magnitude of the relationship among judgments of voice quality and the significant predictors.

Table 3.13 Effect Size: MANOVA Results

<table>
<thead>
<tr>
<th>Effect</th>
<th>Partial Eta-Squared</th>
<th>p-value</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group</td>
<td>0.040</td>
<td>0.000</td>
<td>Small</td>
</tr>
<tr>
<td>Stimuli</td>
<td>0.120</td>
<td>0.000</td>
<td>Medium</td>
</tr>
<tr>
<td>Interaction</td>
<td>0.006</td>
<td>0.000</td>
<td>Small</td>
</tr>
</tbody>
</table>

Table 3.14 Effect Size: ANOVA Results

<table>
<thead>
<tr>
<th>Rating</th>
<th>Effect</th>
<th>Eta-Squared</th>
<th>p-value</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Severity</td>
<td>Group</td>
<td>0.05</td>
<td>0.000</td>
<td>Small</td>
</tr>
<tr>
<td></td>
<td>Stimuli</td>
<td>0.05</td>
<td>0.000</td>
<td>Small</td>
</tr>
<tr>
<td>Roughness</td>
<td>Group</td>
<td>0.05</td>
<td>0.000</td>
<td>Small</td>
</tr>
<tr>
<td></td>
<td>Stimuli</td>
<td>0.20</td>
<td>0.000</td>
<td>Large</td>
</tr>
<tr>
<td>Breathiness</td>
<td>Group</td>
<td>0.03</td>
<td>0.000</td>
<td>Small</td>
</tr>
<tr>
<td></td>
<td>Stimuli</td>
<td>0.06</td>
<td>0.000</td>
<td>Medium</td>
</tr>
<tr>
<td>Strain</td>
<td>Group</td>
<td>0.09</td>
<td>0.000</td>
<td>Medium</td>
</tr>
<tr>
<td></td>
<td>Stimuli</td>
<td>0.07</td>
<td>0.000</td>
<td>Medium</td>
</tr>
<tr>
<td>Pitch</td>
<td>Group</td>
<td>0.11</td>
<td>0.000</td>
<td>Medium</td>
</tr>
<tr>
<td></td>
<td>Stimuli</td>
<td>0.03</td>
<td>0.000</td>
<td>Small</td>
</tr>
<tr>
<td></td>
<td>Interaction</td>
<td>0.02</td>
<td>0.000</td>
<td>Small</td>
</tr>
</tbody>
</table>
3.2 Experiment Two

Results of experiment two were analyzed using INDSCAL, an individual scaling system for MDS (Caroll & Chang, 1970). INDSCAL generated a group space calculated by the angle of separation between vectors for individual responses. In this experiment, results were based completely on perceptions of each group. INDSCAL was chosen from the ALSCAL algorithm in SPSS (version 17; SPSS Inc., Chicago, IL, USA). INDSCAL solutions were obtained in two through six dimensions. The values of stress and the amount of variance accounted for each solution are shown in Figure 3.11. Based on these results, a three dimensional solution, accounting for 92.67% of the variance, was chosen because the increase in variance accounted for by adding a fourth dimension was very small. Dimension one accounted for 42.57% of the variance, dimension two accounted for 39.61% of the variance, and dimension three accounted for 10.47% of the variance. Iterations were stopped at four because the s-stress improvement was less than 0.001. (See Table 3.15) Individual stress levels and $R^2$ values for each group matrix are outlined in Table 3.16.
Figure 3.11 Scree Plot

![Scree Plot](image)

Table 3.15 S-Stress Improvement

<table>
<thead>
<tr>
<th>Iteration</th>
<th>S-Stress</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.17215</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.16331</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.14636</td>
<td>0.01695</td>
</tr>
<tr>
<td>3</td>
<td>0.14496</td>
<td>0.00140</td>
</tr>
<tr>
<td>4</td>
<td>0.14438</td>
<td>0.00058</td>
</tr>
</tbody>
</table>

Table 3.16 Group Stress Levels and $R^2$ Values

<table>
<thead>
<tr>
<th>Matrix</th>
<th>Stress</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (SLPs)</td>
<td>0.148</td>
<td>0.913</td>
</tr>
<tr>
<td>2 (SLPGRADs)</td>
<td>0.138</td>
<td>0.922</td>
</tr>
<tr>
<td>3 (SLPGRADV)</td>
<td>0.144</td>
<td>0.923</td>
</tr>
<tr>
<td>4 (SVTs)</td>
<td>0.111</td>
<td>0.949</td>
</tr>
<tr>
<td>5 (SVTGRADs)</td>
<td>0.131</td>
<td>0.936</td>
</tr>
<tr>
<td>6 (IEs)</td>
<td>0.135</td>
<td>0.917</td>
</tr>
</tbody>
</table>

Dimensions were determined by calculating Pearson $r$ correlations, tested against zero at an alpha level of 0.01, between stimulus coordinates and acoustical measures. A relationship
was considered significant at an alpha level of 0.01 and when the correlation accounted for fifty or more percent of the variance. D1 was correlated with jitter \((r = -0.705, p<0.00)\) and shimmer \((r = -0.708, p<0.00)\). These are both perturbation measures, indicating variability in the signal, and were the bases of listener perceptions for dimension one. D2 correlated with NHR \((r = 0.701, p<0.00)\), Pearson \(r\) at autocorrelation peak (RPK) \((r = -0.706, p<0.00)\), and vocal turbulence index (VTI)\((r = 0.738, p<0.00)\). These are all additive noise measures indicating that strength of voicing was the basis of listener perceptions for dimension two. D3 did not correlate to any acoustical measures and remains unknown.

Table 3.17 reports the weirdness and individual weights for each group. Weirdness values close to zero indicate that the participant has dimension weights proportional to the average (Kreiman et al., 1994). Group weights report the importance of each dimension to each group. D1 (perturbation measures) was weighted heavily for SVTGRADs and D2 (additive noise) was weighted heavily for SLPs. IEs weighed D1 (perturbation measures) and D2 (additive noise) the same. D3, undetermined, was not weighed heavily for any one group above D1 and D2.

<table>
<thead>
<tr>
<th>Group</th>
<th>Weirdness</th>
<th>Weight D1</th>
<th>Weight D2</th>
<th>Weight D3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1(SLPs)</td>
<td>0.3798</td>
<td>0.30</td>
<td>0.82</td>
<td>0.38</td>
</tr>
<tr>
<td>2(SLPGRADs)</td>
<td>0.0961</td>
<td>0.58</td>
<td>0.68</td>
<td>0.35</td>
</tr>
<tr>
<td>3(SLPGRADVs)</td>
<td>0.2418</td>
<td>0.74</td>
<td>0.59</td>
<td>0.20</td>
</tr>
<tr>
<td>4(SVTs)</td>
<td>0.0466</td>
<td>0.63</td>
<td>0.74</td>
<td>0.31</td>
</tr>
<tr>
<td>5(SVTGRADs)</td>
<td>0.5546</td>
<td>0.91</td>
<td>0.28</td>
<td>0.14</td>
</tr>
<tr>
<td>6(IEs)</td>
<td>0.2198</td>
<td>0.60</td>
<td>0.60</td>
<td>0.45</td>
</tr>
<tr>
<td>Average</td>
<td>0.43</td>
<td>0.40</td>
<td>0.10</td>
<td></td>
</tr>
</tbody>
</table>

A 2D representation of each dimension shows a partial separation among groups. Figure 3.12 indicates that SLPs and SVTGRADs are significantly different from the other groups across the first and second dimensions with SVTs, IEs, and both SLP graduate student groups clustering...
together. SLPs weighed additive noise more than any other group and perturbation measures less than any other group. SVTGRADs were the opposite; weighing perturbation measures more than any other group and additive noise less than any other group. All other groups remained more toward the center of the plot, weighing perturbation and additive noise measures similarly. Figure 3.13 shows that dimension three had different grouping with SVTGRADs, SLPGRADVs, and IEs separating from the other groups and SLPs, SLPGRADs, and SVTs clustering together. IEs weighed dimension three heavily while SVTGRADs and SLPGRADVs, similarly, did not weight dimension three heavily across their judgments of similarity.
Figure 3.12 Groups plotted as D1 Vs. D2

Note. 1 – SLPs, 2 – SLPGRADs, 3 – SLPGRADVs, 4 – SVTs, 5 – SVTGRADs, and 6 – IEs.
A more significant separation of groups across all three dimensions can be seen in Figure 3.12. Although SVTs and SLPGRADs are grouped together, SLPs, SLPGRADVs, SVTGRADs, and IEs are separated from this small cluster and one another. These results show a difference in perceptual spaces being used among SLPs, SLP graduate student groups, SVTGRADs, and IEs. Although SLP graduate students who had not had a voice client and SVTs are grouped similarly across three dimensions, these two groups separated from all others when judging similarity. Again, none of the groups weighed the third dimension more heavily than D1 or D2.
Figure 3.14 3D Plot of Group Weights

Note. 1 – SLPs, 2 – SLPGRADs, 3 – SLPGRADVs, 4 – SVTs, 5 – SVTGRADs, and 6 – IEs.
4. CHAPTER FOUR - DISCUSSION

4.1. Experiment One

Experiment One examined the effect of level and type of experience when rating voice qualities with systematically altered acoustical measurements using a standardized rating scale. Research has indicated that internal standards (experience) affect perceptions of voice quality (Kreiman et al., 1993; Kreiman et al., 2004; Kreiman et al., 2007) under the premise that if listeners process the acoustic signal similarly because of a shared background and/or experience, then their ratings of voice quality will also be comparable (Kreiman & Gerratt, 1996). Group and type of stimulus were consistent significant factors for each scale implying that experience as well as jitter, shimmer, and NHR had an effect on ratings of overall severity, roughness, breathiness, strain, and pitch. Most importantly, type of experience seemed to be more of a differentiating factor than level of experience.

4.1.1. Intrarater Agreement

Singing voice teachers were the only group who did not agree with themselves from rating one to rating two, specifically for ratings of overall severity and roughness. Tenured singing voice faculty may be using an inconsistent basis of perceptual measurement to rate overall severity and roughness, because these are two scales that are foreign to their everyday practice unlike the voice qualities of breathiness, strain, and pitch. Overall severity and roughness are only related to dysphonia whereas breathiness, strain, and pitch qualities may be present while training elite voices.

4.1.2. Interrater Agreement

A difference in agreement was evidenced by Sofranko and Prosek (2012) among speech-language pathologists, singing voice teachers, and inexperienced listeners. Expert listeners
demonstrated a better interrater agreement than inexperienced listeners. However, Sofranko and Prosek (2012) examined the agreement among listeners when making categorical judgments. In contrast, when using continuous scales, Experiment One showed that expert listeners were found to have a better (moderate) interrater agreement as compared to student groups and inexperienced listeners only for ratings of overall severity and roughness. SLP student groups had higher levels of interrater agreement for ratings of breathiness over expert groups. Voice qualities of overall severity and roughness are more obscurely defined as compared to breathiness which can explain why students may have obtained a higher agreement. Students may have less difficulty identifying breathiness and its definition throughout their coursework as compared to the other voice qualities. Only ratings of overall severity, roughness, and breathiness had significant correlations amongst most groups. These voice qualities correspond to the acoustical measures systematically altered within the signals; therefore, jitter, shimmer, and NHR are important to perceptual judgments as listeners tracked these changes. Muscular effort (strain), as well as fundamental frequency (pitch), were not directly manipulated. Low agreement for ratings of strain and pitch among experienced and inexperienced listeners is not uncommon as this was evidenced by Helou, Solomon, Henry, Coppit, Howard, and Stojadinovic (2010). Although, interrater agreement levels for the CAPE-V were slightly higher for ratings of overall severity as compared to this study, listeners were provided with training prior to the experiment as well as anchors representing moderate overall severity during the experiment. Moderate correlations among expert listeners in this study may be due to the fabrication of independent baselines for judgments of voice quality secondary to the label of “expert.” Expert listeners may be forming their own definitions for overall severity and roughness. If listeners
are using their own definitions for common voice qualities, updated definitions may be necessary to appropriately re-examine agreement levels during perceptual tasks.

4.1.3. Experience

In Experiment One, inexperienced listeners did not significantly or consistently differ from the other groups when rating overall severity, roughness, breathiness, strain, and pitch. Throughout the study, IEs were at times grouped with the individuals with a singing background, at times grouped with the individuals with a speech-language pathology background, or a combination of both. Their ratings varied in agreement between these two different types of experience. This lack of consistency for judgments of voice quality may be indicative of their lack of experience or common academic coursework to obtain consistent grouping with other listeners when judging voice quality. According to mean ratings, individuals with a singing background demonstrated a sensitivity to disordered voice quality whereas individuals with a speech-language pathology background demonstrated an insensitivity to disordered voice quality. In four out of five scales, IEs were in between these two types of experience with neither a sensitivity nor insensitivity to the stimuli. Also, findings may be due to a difference in rating task as Sofranko and Prosek (2012) used categorization as compared to this study’s continuous rating scale.

The student groups were not significantly different from their corresponding expert groups across the five rating scales. A more consistent difference was demonstrated between individuals with a speech-language pathology background and those with a singing voice background, including the student groups. Very often, participants with a singing voice background differed from other groups (speech-language pathologists, speech-language pathology graduate students groups, and inexperienced listeners) across most of the rating scales.
As mentioned, for all quality rating scales, individuals with a singing background rated the stimuli more severely than the speech-language pathology groups. In four out of five rating scales (overall severity, roughness, breathiness, and strain), SVTs demonstrated the highest ratings of severity. This may be due to a difference in skill set when listening to voice qualities, particularly a sensitivity to disordered voice. SVTs, and their students, spend much of their day listening to normal or elite voice quality, specifically aesthetics. Abnormal voice quality may seem more deviant as compared to groups who have experience with dysphonia. Although mean ratings were not always significantly different from all other groups, SVTS, SVTGRADs, and IEs consistently demonstrated (in varying order) the highest ratings of severity across all five scales. This is also consistent with results from Helou et al. (2010) in which inexperienced listeners in the area of dysphonia rated voice qualities more severely. SLPs and SLP graduate student groups (in varying order) always demonstrated the lowest ratings of severity across all scales which again may indicate that experience with voice disorders will desensitize one’s perception of dysphonia.

Only ratings of overall severity and roughness had significantly different means per type of experience. Ratings for breathiness, strain, and pitch had significant overlap among group ratings, except at extreme ends of mild and severe averages, because perhaps these voice qualities are more easily agreed upon regardless of type and level of experience. The voice qualities of overall severity and roughness are more specific to individuals with a speech-language pathology background as compared to the other groups.

4.1.4. Stimuli

Previous research studies have shown that there are inconsistent relationships among acoustical measurements and perceptions of voice quality (Bhuta et al., 2004; Deal &
Emmanuel, 1978; Eskenzai, et al., 1990; Hecker & Kruel, 1970; Lieberman, 1963; Martin et al., 1995; Parsa & Jamieson, 2001; Prosek et al., 1987; Smith et al., 1978; Wolfe & Martin, 1997; Wolfe & Steinfatt, 1987; Yumoto et al., 1982; Yumoto et al., 1984) In contrast, across all groups in this study, jitter and shimmer combination stimuli and shimmer only stimuli had a consistent relationship with the highest ratings of overall severity, roughness, and strain (with the exception of SVTGRADs with jitter only stimuli having higher means than shimmer only stimuli on ratings of strain) regardless of level and type of experience. Consistently, NHR stimuli had the lowest mean ratings and one can conclude that additive noise had little contribution for ratings of overall severity, roughness, and strain.

More consistent with previous literature (Colton & Casper, 1996; Hillenbrand, 1988; Hillenbrand & Houde, 1996), NHR stimuli had a consistent significant correlation with ratings of breathiness as well as the highest mean ratings of breathiness across all groups. Breathiness ratings were the only scale in which NHR stimuli did not have the lowest means among ratings. In addition, all other manipulated acoustical measures had insignificant correlations with ratings of breathiness. Therefore, alterations of NHR will reliably lead to ratings of breathiness across all levels and types of experience.

Following a similar pattern for overall severity, roughness, and strain, jitter/shimmer combination stimuli had the highest mean ratings with NHR having the lowest mean ratings for pitch. Although patterns were slightly inconsistent, jitter and shimmer combination stimuli, jitter only stimuli, and shimmer only stimuli, were contributors to perceptual judgments of pitch. However, it is important to note that only SVTs, SVTGRADs, and IEs had significant correlations with altered acoustical measurements for these ratings. Although fundamental frequency was held constant for all stimuli, individuals with a singing background frequently
commented on deviant pitch patterns throughout the study. Pitch is a common perceptual category that individuals with a singing background use throughout their professional careers. A special sensitivity to tone quality may be the cause for significant correlation to pitch judgments. In contrast to this theory, IEs also demonstrated significant correlations between acoustical measures and pitch ratings. However, as stated earlier, IEs were alternating between ratings consistent with individuals with a speech-language pathology background and those with a singing background leading to chance in judgment rather than a steady baseline for perceptual ratings.

4.2. Experiment Two

Experiment Two examined the perceptual spaces of the listeners with varying levels and types of experience when judging the similarity among pairs of stimuli with systematically altered acoustical measurements. Previous literature discusses the multidimensional nature of voice quality (Bassich & Ludlow, 1986; Kempster et al., 1991; Kreiman et al., 2004; Kreiman et al., 2007; Murry, Singh, & Sargent, 1977). A three dimensional configuration was chosen and, when plotted, showed that the groups did separate from one another (with the exception of SVTs and SLPGRADs grouping more closely together) indicating that individuals with different types and levels of experience are using different perceptual spaces when judging similarity of stimulus pairs with systematically altered acoustical measurements. Most importantly, SLPs, who treat disordered voice quality, consistently separated from all other groups for D1 (perturbation) and D2 (additive noise).

Dimension one correlated with jitter and shimmer, more generally stated, perturbation measures or signal instability. These two perturbation measures were systematically altered to generate the stimuli and were tracked by all listener groups. Perturbation measures have been
shown to be highly correlated with dysphonic voice quality (Bhuta, Patrick, & Garnett, 2004; Deal & Emmanuel, 1978; Eskenzai, et al., 1990; Hecker & Kruehl, 1970; Lieberman, 1963; Hillenbrand, 1988; Martin et al., 1995; Smith et al., 1978; Wolfe & Martin, 1997; Wolfe & Steinfatt, 1987; Yumoto, Sasaki, & Okamura, 1984). SLPs weighed this dimension less than any other group. This may be due to an overexposure to instability in voice quality, as compared to any other group, secondary to treating dysphonic voices. SVTGRAD students, weighing this dimension more heavily than the other groups, may be overly sensitive to instability in voice as they are just beginning their focus on voice quality perception through a singing pedagogy course. Voice qualities that come as a result of increased perturbation measures may be a stark contrast to their new academic coursework and new experiences. SLP graduate student groups and SVTs have had some exposure, but not enough to parallel that of SLPs, leaving them to weigh D1 in the midrange. IEs may have weighed D1 also in the midrange secondary to uncertainty regarding the information embedded within the signal rather than a similarity with more experienced groups.

Dimension two correlated with NHR, Pearson $r$ at autocorrelation peak (RPK), and vocal turbulence index (VTI). NHR, a common measure for breathiness, was systematically altered to generate the stimuli. RPK, based on the fundamental period, is a correlation of the signal and a delayed version of itself (Hillenbrand et al., 1994). Periodic signals (normal voices) will have more prominent peaks for correlation than breathy signals, leading to weaker correlations for less prominent signals (breathiness) (Hillenbrand et al., 1994). Lastly, VTI is a measure of turbulence caused by incomplete glottal closure during phonation, again breathiness (Xue & Deliyski, 2001). All of these measures are related to strength of voicing/additive noise. The results show that SLPs weighed this dimension more heavily than any other group. This is an
indication of a particular sensitivity or ability to easily detect to strength of voicing that comes after three or more years of experience with dysphonia. SVTGRAD students, again opposite of SLPs, did not weigh this dimension heavily, separating from all other groups. Perhaps these students, just learning to focus on the balance between airflow and phonation, make allowances for additive noise in regard to similarities between stimuli. Also, SVTGRADs may not view breathiness as an extreme voice quality secondary to their extensive focus on breathing technique and breath support in their coursework. All other groups (IEs, SVTS, and both SLP graduate students groups) demonstrated an equal distribution of weight between additive noise and variability in voice quality when rating pairs of stimuli. Again, SVTs, and SLPGRAD student groups have had exposure to noisy signals but not enough to parallel SLPs. Only IEs weighed both dimension one and two exactly the same, and as stated, this may be due to uncertainty rather than a comparison with the more experienced groups.

All systematically altered measures (perturbation and additive noise) were accounted for in D1 and D2. Listeners consistently tracked these changes when judging similarity. Perturbation and additive noise were the only altered components of the generated stimuli and as a result, dimension three did not correlate to any acoustical measures. This indicates that dimension three is unrelated to acoustical measures. Dimension three may be a representation of a top-down effect where listeners are bringing their own components to the task rather than a manipulation of the signal affecting judgment (bottom-up effect). Lindblom (1990, p. 228) explains that “Our perception of speech and other communicative events is not determined by the signal alone. It is shaped by an interaction between the signal on the one hand and information stored in our brains on the other.” Dimension three, unrelated to the signal, may be related to factors each participant brought to the perceptual rating task.
4.3. Limitations and Future Research Suggestions

There were a few limitations to the study that should be noted. Participants reported the absence of a hearing impairment, but hearing thresholds were not obtained prior to the study. Although participants were able to adjust the playback volume to a comfortable level, this is of concern as many participants were over the age of 30 and hearing thresholds may have been diminished. Considering the possibility of advanced ages when including expert populations, future research should include hearing screenings prior to listening tasks to assure hearing thresholds are within normal limits.

Voice quality definitions may have been another limitation to the study. The definitions of voice qualities provided during the study were cause for some voiced confusion for groups without a speech-language pathology background. This confusion may have impacted the judgments of the listeners and ultimately the results of the study. In addition, the word “strain” was used in the roughness definition. Results show that judgments of strain and roughness had correlations with the same acoustical measures. It may not be that these two voice qualities are analogous, but rather an effect of similarity in definition. Also, several speech-language pathologists reported they currently use voice quality definitions including more physiological terms. Future research studies need to consider the interpretation of provided definitions or conduct a study to obtain new definitions, more agreed upon by the expert population, prior to experimentation.

One approach to this latter experiment would be to have the experts suggest definitions for each of the voice qualities typically used in clinical practice. No restrictions need to be placed on the definitions, and the experts would be free to use acoustical, physiological, or any combination of descriptors that were meaningful to them. The definitions could then be
submitted to an expert panel similar to the one described by Kempster et al. (2009) to determine the adequacy and preference of each definition. Definitions obtaining the highest agreement among the panel members could be offered to the profession as consensus definitions to be adopted by voice therapists.

Lastly, group selection may have impacted the results, specifically interrater agreement levels. Some of the speech-language pathologists in the study treated singers, while some did not. These SLPs may have had additional experiences closer to that of an individual with a singing voice background. Also, approximately half of the SVTs taught classical singing while the other half taught musical theater singing. This difference in style may impact judgments of voice quality ultimately decreasing agreement. Future research studies may want to consider these differences in experience within groups when determining inclusion criteria.
Overall, both experiment one and experiment two examined the differences in perceptions of voice quality. Both experiments were conducted using systematically altered acoustical measurements. Differences were examined between experienced and inexperienced listeners and the relationship between perceptions of voice quality and systematically altered acoustical measurements was explored. However, only experiment two examined the differences in perceptual spaces between experienced and inexperienced listeners and the dimensions representing acoustical measurements correlated with similarity judgments.

Level and type of experience may not affect agreement levels with the use of continuous rating scales and synthesized sustained vowel stimuli. Although only the experienced listeners demonstrated higher correlations for overall severity and roughness, these higher correlations were moderate at best, agreeing with previous literature regarding listener agreement. Expert listeners may have developed their own unique definitions of each voice quality. Definitions of common voice qualities may need to be updated to better represent the expert listeners working in the area of voice currently. Also, academic programs may not be preparing students to adequately identify common dysphonic voice qualities.

There was not a significant difference in ratings of perceptions of voice quality between individuals with different levels of experience when systematically altering acoustical measurements. However, a consistent difference in judgment between types of experience (individuals with a singing background and individuals with a speech-language pathology background) was observed. Inexperienced listeners followed an inconsistent pattern, never being significantly different from either types of experience. They were frequently placed in the middle demonstrated neither a sensitivity nor insensitivity to the stimuli.
Although type of experience yielded different results for ratings of overall severity, roughness, breathiness, strain, and pitch, there was some agreement across all groups regarding correlations between altered acoustical measurements and voice quality perceptions. Jitter/shimmer combination stimuli and shimmer only stimuli contributed to ratings of overall severity, roughness, and strain, the same stimuli with the addition of jitter only stimuli contributed to ratings of pitch for three of the six groups. Lastly, NHR ratio stimuli contributed to ratings of breathiness.

Overall, listeners may be using the same acoustical component of the signal for standardized rating scales, but severity ratings will be largely affected by type of experience. Individuals with experience in disordered voice quality tend to give lower ratings of overall severity, roughness, breathiness, strain, and pitch. Individuals with experience specifically with the aesthetics of normal or elite voice quality tend to give higher ratings of overall severity, roughness, breathiness, strain, and pitch. Individuals without a specified type of experience varied in regard to having similarities with a background in singing, a background in speech-language pathology, or both.

Within Experiment Two, any bias created by the rating scale was removed and the question of quality definitions was not a factor. The listeners were only asked to rate similarity among pairs of stimuli and dimensions were determined by the perceptions of each group and not the scale. Listeners were using perceptual dimensions of perturbation and additive noise when judging similarity. Only changes in perturbation (jitter and shimmer) and additive noise (NHR) were used to generate the stimuli and these measures were accounted for in the first two dimensions. Dimension three, although significant, remains unknown. This dimension is a non-experimental dimension unrelated to any manipulations of the signal but rather what the
individual has brought to the perceptual rating task which can be any number of factors. For example, dimension three could be related to physiology, attention, memory, etc.

Although Experiment One did not show a difference between judgments for different levels of experience, Experiment Two indicated that groups with different types and levels of experience were using three dimensions to rate similarities, but in different ways. Using a 3D representation, this separation was evident. Group differences were less evident with 2D representations, but a consistent separation with SLPs and SVTGRADs occurred for both D1 and D2.

SLPs are the professionals responsible for the treatment of dysphonia because they have extensive training, academic coursework, and experience with disordered voice quality; however, SLP graduate students are often used as listeners for research studies as they have had a course on voice disorders secondary to being readily available. This study did not directly examine accuracy of judgments of voice quality but rather a possible difference in judgments of voice quality with and without a commonly used rating scale. The results of Experiment One showed a difference in type of experience when using a commonly used rating scale and raise questions regarding the use of this scale based on the obtained agreement levels. However, Experiment Two gives evidence to the importance of using SLPs with experience as judges, rather than SLPGRAD students, to better generalize results of perceptual studies to the field. A consistent difference in perceptual spaces can be seen between SLPs and SLP graduate students (who have had a voice client and those who have not had a voice client). Also, although individuals with a background in singing voice have extensive experience with perceptions of voice quality, the study reveals that their experience does not equate to that of individuals with a background in speech-language pathology.
Overall, this study agrees with Sofranko and Prosek (2012) in that experience can affect judgments of voice quality. Individuals with different levels and types of experience will rate voice differently when given common voice quality definitions and will use different perceptual spaces when rating similarity. By controlling many of the factors that affect perceptions (experience, rating scale, and stimulus type/length), a consistent relationship between subjective and objective measures was revealed. The common acoustical measures of jitter, shimmer, and NHR, will reliably lead to differences in perceptions of voice quality.

This study shows that speech-language pathologists with experience in dysphonia differ in ratings as compared to individuals with other types of experience and do not share a perceptual space with SLP graduate students. The continued use of SLP graduate students as well as groups with mixed types of experience in research studies may not be adequate to examine relationships among subjective and objective measures of voice. Due to continued moderate agreement levels as well as an unidentified dimension accounting for perceptions of voice, continued research is needed in the area of voice perception using clearly defined groups. With continued control of the factors that can affect perceptions of voice quality, researchers may continue to reveal consistency between subjective and objective measures of voice to benefit clinicians and patients.
References


Appendix A – Instructions for Experiment One

You will be presented auditorily with a series of voice samples. For each sample you will be asked to rate the voice quality on a continuous scale ranging from mild to severe (MI = mild, MO = moderate, SE = severe) for five different perception categories: "overall severity," "breathiness," "roughness," "strain," and “pitch adequacy.”

Overall severity is defined as "a global, integrated impression of voice deviance" (CAPE-v, Kempster et al., 2009, p.130).

Breathiness is defined as an escape of air during voicing that one can hear which reduces loudness (Colton & Casper, 1996).

Roughness is defined as a voice quality lacking clarity with increased noisiness characterized by tension or strain one can hear (Colton & Casper, 1996).

Strain is defined as difficulty maintaining and controlling voicing with moments of voice loss (Colton & Casper, 1996).

Pitch adequacy is defined as a “pitch that deviates from normal for that person’s gender, age, and referent culture” (CAPE-V, Kempster et al., 2009, p.130) The individual you are listening to a 23 year old Caucasian female.

Each stimulus is approximately 1 second of the sustained vowel /a/ and this experiment includes a total of 70 stimuli.

You may replay the stimuli at any time during this experiment.

Once you click the button labeled "okay" your ratings will be submitted and you will proceed with the next stimulus. However, you may back up to the previous stimulus if you desire by pressing the button labeled "back up."

You may terminate this experiment at any time by indicating so to the experimenter. When you are ready to begin, click "Start."
Appendix B – Instructions for Experiment Two

You will be presented auditorily with a series of paired voice samples. For each pair, you will be asked to rate the similarity between them on a scale ranging from "no similarity" to "extremely similar" using a slider scale. You may click on any point on the line to indicate your rating. You cannot replay the stimuli during the experiment and you will not be able to return to previous stimuli. Once you click the button labeled "okay," your rating will be submitted and you will proceed with the next pair of stimuli. You may terminate this experiment at any time by indicating so to the experimenter. When you are ready to begin, click "Start."
VITA
Jessica Lynn Sofranko

Degrees Awarded:
2012 Ph.D. Communication Sciences and Disorders, The Pennsylvania State University
2007 M.S. Speech-Language Pathology, Misericordia University
2007 B.S. Psychology, Misericordia University

University Teaching Experience
2011-2012 Instructor, Department of Communication Sciences and Disorders, The Pennsylvania State University
2011 Clinical Supervisor, Department of Communication Sciences and Disorders, The Pennsylvania State University
2008-2009 Clinical Supervisor, Speech-Language Pathology, Misericordia University

Professional Experience
2007-2009 Speech-Language Pathologist, Janet Weis Children’s Hospital, Pediatric Rehabilitation, Geisinger Medical Center, Danville, PA
2010-present Per Diem Speech-Language Pathologist, Greater Hazleton Health Alliance, Hazleton, PA
2010-present Per Diem Speech-Language Pathologist, Ambassador Rehabilitation, The Village at St. Luke’s, Hazleton, PA
2010-present Per Diem Speech-Language Pathologist, Guardian Rehabilitation, Elder Care and Rehabilitation Center, Mountain Top, PA.

Publications

Presentations
Sofranko, J.L., Tellis, C.M., Clarke, K., & Tellis, G.M. (2010). Pediatric caregivers’ perception of the effectiveness of neuromuscular electrical stimulation. ASHA Convention, Philadelphia, PA.
Tellis, G.M., Sofranko, J., Frey, K., Bressler, L., & Kyle, K. (2009). Middle and high school students views about bullying and stuttering. ASHA Convention, New Orleans, LA.

Awards and Honors
2007 Academic Excellence Award in Speech-Pathology, Misericordia University
2010 MaryAnn Peins Graduate Scholarship in Speech Pathology
2010 ASHA Award for Continuing Education