A COMPARISON OF PSYCHOPHYSICAL METHODS FOR THE EVALUATION OF ROUGH VOICE QUALITY

By

STACIE NOELLE CUMMINGS

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

UNIVERSITY OF FLORIDA

2010

© 2010 Stacie Noelle Cummings

ACKNOWLEDGMENTS

This thesis would not have been even remotely possible without the guidance and encouragement of my committee chair, Dr. Rahul Shrivastav. His enthusiasm for research and willingness to teach is the reason I decided on pursuing a thesis. I would also like to thank Dr. Judith Wingate for her encouragement and support throughout both my thesis and my graduate clinical experiences. Many thanks to Dr. Sona Patel, for her patience in teaching me proper data collection procedures, staying up all hours of the night to practice presentation delivery, and explaining the answers to the million questions I am sure I asked her. Thanks to Dr. David Eddins for his collaboration in all our research and for all his encouragement. A special thanks to Evan McIntyre, for his endless support of my developing coffee habit; to Jason Sanders, for sending me a constant stream of information about graduate thesis guidelines and procedures; and to all of my families who have cheered me to the finish. Finally, I would like to thank the NIH R01 grant for funding the research on voice quality.

TABLE OF CONTENTS

page

LIST OF TABLES

LIST OF FIGURES

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

A COMPARISON OF PSYCHOPHYSICAL METHODS FOR THE EVALUATION OF ROUGH VOICE QUALITY

By

Stacie Noelle Cummings

May 2010

Chair: Rahul Shrivastav
Major: Communication Sciences and Disorders

Previous research has used rating scales, rankings, and matching tasks to quantify roughness in voice quality. Each method has drawbacks, such as the variability of rating scale data, the nature of ordinal data obtained through rankings, and the lengthiness of matching tasks. However, the method which has produced the most reliable and measurable data is the matching task. In the present experiment, a method of constant stimuli task was used to quantify roughness in dysphonic voices and analyzed in comparison with the findings reported from other tasks.

Ten participants listened to vowels produced by 34 disordered voices with varying degrees of roughness. These were paired with a synthesized signal in which the modulation depth of the waveform amplitude was systematically varied. For each pair the participant chose which ‘voice’ had a greater amount of roughness. The data obtained through the method of constant stimuli task was compared to that obtained for the same 34 disordered voices from the rating scale, ranking, and the adaptive matching tasks.

The results indicate that similar perceptual distances are gathered using each of the methods; however, the adaptive matching task appears to be optimal for evaluating voice quality in a relatively time-efficient, accurate and quantifiable manner.

CHAPTER 1
INTRODUCTION

Voice quality is a perceptual phenomenon. It can be defined as a “term that includes all the leftover perceptions after pitch, loudness, and phonetic category have been identified” (Titze, 1994). These perceptions include quality descriptors such as breathiness, roughness, nasality, register, and strain. Since the ‘quality’ of a voice is determined by how the listener hears and judges the sound, finding an accurate and reliable means by which a particular voice quality can be measured is essential. Traditionally, in the field of speech-language pathology, clinical perceptual judgments of voice quality are obtained using rating scales. These ratings are generally used by clinicians to establish a baseline, to indicate a level of disordered voice quality in referrals and reports, and to serve as evidence for pre- and post-treatment outcomes.

However, a more accurate method of quantifying voice quality is necessary in order to understand how listeners perceive voice quality and to develop objective measures of voice quality that can enable more accurate and consistent measures across clinicians or therapy sessions, and which may be used in a wide range of research and technology applications. Some of the far reaching applications may include creating more realistic synthesized voices, aid in voice recognition technology, detection of indexical properties in speech such as emotions and to assist in the early detection and diagnoses of various clinical conditions through examination of the speech acoustic signal.

Roughness is a voice quality that may be defined as the “psycho-acoustic impression of the irregularity of vocal fold vibrations […] which corresponds to the irregular fluctuations in the fundamental frequency and/or the amplitude of the glottal

source sound" (Hirano, 1981). Many different acoustic correlates have been associated with the percept of roughness. A 1990 study by Eskanazi proposed that Spectral Flatness of the Residue Signal (SFR) and Harmonics-to-Noise Ration (HNR) were the acoustic correlates which were most closely related to roughness. Other studies have suggested percent shimmer or jitter, amplitude variability index, or normalized noise energy as correlates for roughness (Heiberger & Horii, 1982; Deal & Emmanuel, 1978; Hirano et al., 1988); however, few have proven to be strong indicators of rough voice quality across different experiments and stimuli. Another difficulty encountered in the measurement of rough voice quality is in isolating the percept of roughness from others such as breathiness or hoarseness, as they often co-occur. For example, by many definitions, the level of both roughness and breathiness are strongly influenced by noise levels. Thus, the acoustic cues specific to roughness have not yet been clearly determined.

However, Fastl and Zwicker (1999) examined correlates of roughness for simpler, non-speech signals such as pure tones and noises. They determined that roughness was best described by the frequency and depth of the amplitude modulation of the waveform envelope. By systematically manipulating these parameters, they found that the amplitude modulation frequency that resulted in the greatest roughness was dependent upon the frequency of the carrier signal itself. For all carrier signals, roughness increased with an increase in modulation frequency, reached a peak and then reduced. The modulation frequency that resulted in the greatest roughness varied from 25 Hz for a 100 Hz carrier to a maximum of 70 Hz for a 1 kHz carrier tone.

The current experiment is a part of a larger project to understand how listeners perceive and evaluate dysphonic voice quality and to develop methods that can accurately quantify changes in voice quality. A critical element of such work is to obtain precise estimates of voice quality perception. The current study compares three methods of voice quality measurement. The first is the rating scale task, the second is a matching task, and the third uses the method of constant stimuli. In each task, listeners make perceptual judgments of rough voice quality. These judgments are compared in order to determine whether similar perceptual distances can be obtained by each of the methods, and to determine which of the methods is most appropriate for evaluating rough voice quality. Furthermore, data obtained using the method of constant stimuli task may be useful in helping define specific cues associated with the perception of roughness.

CHAPTER 2
LITERATURE REVIEW

Rating Scales

Rating scales are commonly used in research and in clinical settings as a means of collecting perceptual data. In the rating scale, the rater assigns a number to the stimulus which is being evaluated. One example is the pain rating scale typically used in hospitals for patients to self-evaluate and describe the pain they are experiencing (McCaffery & Pasero, 1999). The pain scale ranges from 0 or 'No Pain,' to 10 or 'Worst Possible Pain'. By selecting a number on the scale patients attempt to convey the amount of discomfort they are feeling.

Previous research has indicated that tasks such as rating scales are not always ideal for quantifying voice quality percepts. The problem with such methods lies in the variability inherent in subjective judgments. One study in particular noted that perceptual judgments of one stimulus could be rated on every point of a 7-point rating scale (Kreiman et al., 1993). Factors which increase the likelihood of variability include listener bias, differing levels of experience, and varied exposure to stimuli, amongst others. However, variability is less marked in stimuli which exist on extreme ends of the spectrum; furthermore, variability can be further controlled by averaging and standardizing multiple ratings in an experimental set-up (Shrivastav, Sapienza, & Nandur, 2005).

An additional limitation of rating scales lies in the type of data they yield. Rating scales typically require listeners to choose a number for each stimulus, but fail to offer a means of indicating distance between the numbers or stimuli (e.g., Shrivastav et al., 2005). Thus the listener may be forced to group or separate voices which may not

actually be as perceptually equivalent or different as the assigned number indicates. When a participant assigns a number to a stimulus in this fashion, the type of data is ordinal in nature.

Stevens (1946) details the limitations of ordinal data. According to Stevens (1946), the rules for ordinal scales indicate that they are capable of providing data which shows a relationship of greater than or less than another given point. The mathematical properties of ordinal scales allow for calculation of median scores and percentiles, which have a limited use in quantifying voice quality. However, he notes that other means of analyzing ordinal data, such as means and standard deviations, have been erroneously used for analysis regardless.

Magnitude Estimation

Another method of psychophysical testing is magnitude estimation, which attempts to compensate for the limitations of rating scales. By instead allowing listeners to choose a number in a large range (e.g. 1 to 1000), distances can be more accurately specified. In the magnitude estimation task listeners are instructed to compare the stimulus to a predetermined midpoint, termed the 'modulus', and assign numbers which demonstrate the perceived relationship between the stimulus and modulus (Eadie & Doyle, 2002). For example, if the modulus for 'breathiness' is assigned at 100, the listener might assign 200 to a stimulus perceived to be twice as 'breathy.'

In the case of magnitude estimation, the data obtained is on a ratio scale, which is defined as having units of measurement which indicate equal intervals of distance (Stevens, 1946). This means that the data is more adequately descriptive of the perceptual distances between stimuli. Furthermore, ratio level scales have a zero point which indicates the absence of the property which is being measured. Therefore, the

numbers have a mathematical relationship and statistic functions can be applied. However, the same difficulties with listener bias which are found in rating scales are also evident in magnitude estimation, making it difficult to compare results both across listeners and in repeated tasks.

Adaptive Matching

In order to correct for the variability associated with the arbitrary assignment of numbers to a perceived sound, Patel et al. (2009) suggested using a measurable comparison sound source, or reference signal, as an objective means of quantifying a voice quality. The concept of a reference has been used for many perceptual measures, dating back to Fletcher and Munson's 1933 experiment using a 1 kHz tone to evaluate loudness in other sounds. References have proven useful in many experiments attempting to measure perception of loudness and pitch (for example, Stevens, 1972; Stevens et al., 1937). Data resulting from such experiments have formed the bases for several computational models, such as that for loudness (for example, Moore, Glasberg and Baer, 1997).

Patel et al. (2009) employed an adaptive matching task to collect listener judgments of breathy voice quality. In the adaptive matching task, listeners compare stimuli to a synthesized and measurable reference, which can be systematically varied to match the stimulus. In order to measure breathiness, listeners were allowed to vary the Signal-to-Noise Ratio (SNR) of the reference signal in 1-dB steps until the breathiness of the test voice was perceptually identical to that of the reference signal. This SNR was referred to as the 'threshold' and was used as a measure of breathiness for that voice. The judgments obtained using this method appeared to be systematic and were highly reliable both within and across listeners. The adaptive matching task

was compared with a rating scale and a magnitude estimation task, which confirmed that each method could be used to obtain similar data, but that the matching task was ideal for quantifying breathy voice quality in a manner that facilitated direct comparison across stimuli, listeners or experiments (Patel et al., 2009).

In a follow-up experiment, Shrivastav et al., (2009) provided preliminary data to design an adaptive matching task to measure rough voice quality. Results indicated that listeners were able to use reference signal to judge perceptual distances between rough voice stimuli, but in general found the task to be more difficult than the task for breathiness. In this experiment, listeners used a reference signal consisting of a sawtooth wave and speech-shaped noise as a carrier. The amplitude envelope of this carrier was modulated by a 40 Hz square wave. This frequency was selected based on Fastl and Zwicker's (1999) work on perceived roughness for various amplitude modulated carrier tones; a 40 Hz modulation frequency closely corresponded to peak roughness for pure tones of 100-200 Hz, which is very close to the fundamental frequency of the human voice. The amplitude modulation depth of the reference was manipulated in 2 dB steps by the listener until they perceived that it matched the roughness in the stimulus. The modulation depth at which the listener perceived the reference signal to have the same roughness as the test voice was called the *threshold* and used as a measure of roughness. The listener thresholds demonstrated that listeners were able to judge perceptual distances of roughness using the adaptive matching technique, but that inter-listener reliability was lower than that achieved for breathiness using a similar matching paradigm.

Method of Constant Stimuli

The method of constant stimuli task is a classical method of psychophysical data collection. In this approach, listeners are presented with the stimulus to be evaluated, and with multiple comparison references which have been set to predetermined, fixed levels (Dai, 1995). Upon presentation, listeners are asked to indicate which of two stimuli has more of the property which is being evaluated. The data yielded is in the form of percent correct values, generally ranging from near 0% to slightly under or at 100%. These percent correct values generally take the form of a sigmoid function; thus the data obtained using the method of constant stimuli may also be useful in identifying the perceptual threshold for roughness.

In addition to gathering perceptual data, the current research is driven by the need to understand the cues that specifically lead to the perception of roughness. In order to accomplish this, discovering whether roughness perception involves multiple dimensions is important. One method of doing this is to use a multidimensional scaling technique (MDS). Another means is to determine the psychometric functions for roughness and to see whether different voice stimuli result in similar psychometric functions. Using the method of constant stimuli also allows us to map the psychometric functions for roughness stimuli and helps in this goal.

Summary and Purpose

Of the many methods available for gathering psychometric data on perception of voice quality, three have been chosen for comparison in the present study. The rating scale was chosen due to its widespread use in previous research as well as in clinical settings; however, it is limited by the ordinal data which it produces and the inability to translate measurements from one experiment to another. The adaptive matching task

was selected because of the high reliability and good perceptual distances which it has provided in previous measures of voice quality. The method of constant stimuli was evaluated in order to provide additional perceptual data using a classical psychoacoustic method of measurement. An added benefit of using the method of constant stimuli lies in the fact that it may provide additional information about the cues which are most highly associated with rough voice quality.

The comparison was done in order to (1) validate the newly developed adaptive matching procedure for roughness and (2) to determine the consistency of psychometric functions across multiple voice stimuli. This would help determine the best approach to obtain perceptual judgments of roughness. It may also provide preliminary evidence regarding the acoustic cues for roughness perception. Specifically, the present experiment could help give some insight whether roughness perception in dysphonic voices is affected by cues other than the waveform amplitude modulation alone.

CHAPTER 3
METHODS

Listeners

A total of twenty-seven listeners participated in the study. The mean age of the listeners was 22.8 years; the age range was 18 years to 51 years. Seven of the participants were male, and twenty were female. Listeners were recruited to participate in one of three tasks, a rating scale task, a matching task, or a method of constant stimuli task. Five listeners completed the rating scale task, ten completed the matching task, and ten completed the method of constant stimuli task. Listeners were all native speakers of American English with no previous history of speech problems. Each participant passed a hearing screening bilaterally using air-conduction pure-tone audiometer at 20dB HL for 250Hz, 500Hz, 1000Hz, 2000Hz, and 4000Hz. Participants were recruited from either the department of Communication Sciences and Disorders or the Program in Linguistics at the University of Florida. Listeners for the rating scale task were volunteers. Listeners in the adaptive matching and method of constant stimuli tasks were paid for participating in the experiment.

Stimuli

In each of the tasks, listeners were presented with the same 34 disordered voice samples of the vowel /a/ obtained from the Sataloff/Heman-Ackah voice database. The voices were selected randomly. From each vowel a stable 500 ms sample was extracted and downsampled from 50 kHz to a rate of 12.207 kHz, at 16 bits. Fourteen of the disordered voices were female and twenty were male.

Listeners who participated in the matching task and method of constant stimuli task were also presented with a synthesized signal which was generated using Sykofizx

(Tucker-Davis Technologies Inc., Alachua, FL). The synthesized signal consisted of a low pass filtered 151 Hz sawtooth wave mixed with speech-shaped noise. This synthesized signal is identical to the one described in Patel et al. (2009), and consisted of a 500 ms segment sampled at 12207 Hz with 16 bits quantization. To manipulate the roughness of this signal, the envelope of this wave modulated by a 40 dB raised cosine modulation function. Modulation functions utilizing a sine wave and square wave respectively were attempted in previous pilot work but discarded because they did not provide a good range of roughness. The signal-to-noise ratio was fixed at 12 dB.

Procedures

Listeners were tested in a single-walled sound treated booth. Stimuli were presented in the right ear monaurally at 75 dB SPL using the RP2 processor with ER-2 ear inserts (Tucker-Davis Technologies, Inc., Alachua, Florida). The rating scale task was run using the computer software Ecos/Win (AVAAZ, Inc., Ontario, Canada). The adaptive matching task and method of constant stimuli task were run using the computer software *SykoFizX* (Tucker-Davis Technologies, Inc., Alachua, Florida). Listeners used a keyboard and computer monitor to respond to stimuli.

Rating Scale Task

The rating scale task was a previously conducted, unpublished pilot study on roughness. In the rating scale task, listeners were presented with the 34 rough voice stimuli and instructed to choose a rating for each stimulus on a scale of one to five. On the scale, 1 represented a normal voice, 2 a mildly rough voice, 3 a moderately rough voice, 4 a moderate-severely rough voice, and 5 a severely rough voice. Each stimulus was presented a total of 10 times in a random order. No training paradigm was

implemented for this task. Five listeners participated in the experiment; listeners averaged one 60-minute session to complete testing.

Adaptive Matching Task

The adaptive matching task was also completed in a previous experiment (Shrivastav, 2009). In the adaptive matching task, listeners were presented with each of the 34 stimuli, called the 'standard.' For each standard, listeners also heard a synthesized reference which was termed the 'signal.' The signal was paired with the standard for comparison, and manipulated by the listener to have 'more' or 'less' roughness. This was done when the listener selected either the icon labeled 'more roughness' which increased the modulation depth of the signal by 2 dB, or 'less roughness' decreasing the modulation depth in the signal by 2 dB. When the signal roughness matched that of the standard, listeners indicated that they perceived 'equal' roughness and their answer was recorded; this answer was the listener threshold. For each of the standards, multiple threshold judgments, or 'runs,' were obtained. The starting point from which the listeners heard the signal alternated from a position which was estimated to be more rough and one which would be estimated as less rough than the standard. This approach was used in order to avoid skewing the data based on the starting position. A total of ten runs were collected for each stimulus. This was based on Shrivastav et al. (2008), who showed that increasing the number of runs from a total of six to a total of ten resulted in more consistent data, with the greatest improvements seen for voices with mid- to low-levels of roughness.

In order to prepare listeners for the stimuli presented, a short training session was conducted. The first stage of training, the Interactive Quality Demonstration (IQD), consisted of a GUI displaying five buttons/icons. These represented synthesized voices

that reflected three levels of breathiness, three levels of roughness and two levels of pitch. Breathiness levels were manipulated by varying the signal-to-noise ratio of a base vowel. Roughness was manipulated by modifying the modulation depth of a 40 Hz amplitude modulation of the vowel waveform. Pitch was manipulated by varying the fundamental frequency of the base vowel. Listeners were instructed to attend to differences between varying levels of roughness, between roughness and breathiness, and between roughness and pitch. This was done to help listeners distinguish roughness from breathiness and pitch. After approximately five minutes of familiarization, listeners determined if they felt comfortable with the differences in voice qualities and continued with the next training task.

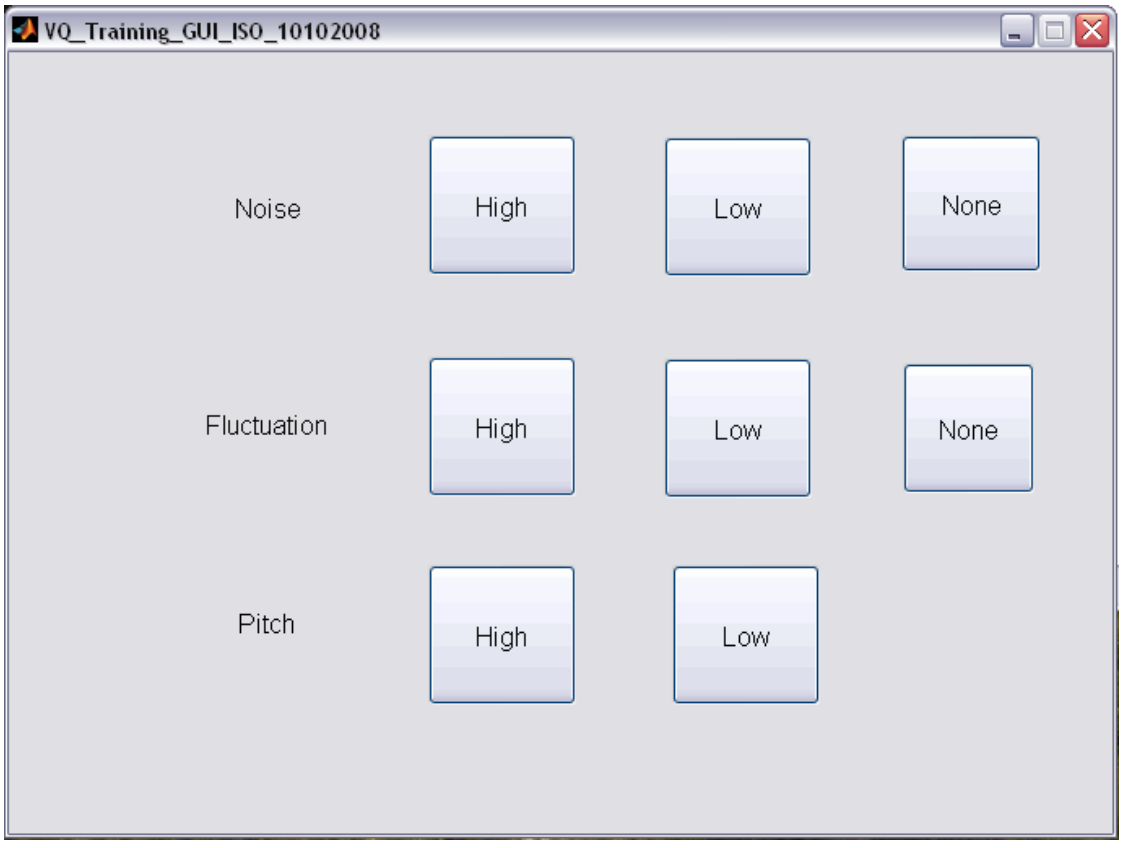

Figure 3-1. The GUI used for the Interactive Quality Demonstration (IQD) training procedure.

Visual Sort Ranking (VSR) was the second training task to be completed by participants. In this task, listeners were instructed to organize and rank order the 34 voice stimuli from ‘least rough’ to ‘most rough’. The GUI allowed listeners to play each voice and re-order the icons as often as needed. When the listener felt the stimuli were in the appropriate order, they selected a button marked ‘Done’. After the initial attempt at sorting stimuli, feedback was provided to the listener. Listener rankings were compared to the consensus rankings given by three expert listeners. If the ranking assigned to a particular stimulus differed from the consensus ranking from expert listeners by more than five, then that stimulus was coded in red (indicating that the listener assigned the icon too much roughness) or green (indicating that the listener assigned the icon too little roughness). If the ranking was one to three units away from the expert consensus, then it was colored cyan but no directional assistance was given for these deviations. If a voice ranking was in agreement with the expert consensus, the icon retained its original grey color. The participants were then allowed to refine their rankings based upon the feedback. On their second attempt at sorting stimuli, a second round of feedback was provided to the listeners; however the voice rankings which were more than five slots were colored magenta regardless of direction of displacement from expert rankings. The rankings which were one to three slots away from expert consensus were cyan. On completion of the sorting for the third time, correlations between the listener rankings and expert rankings were calculated. Only those listeners whose VSR data showed a correlation over 0.70 with the expert listener rankings were permitted to continue in the experiment.

The last phase of training was a matching familiarization task. In this training

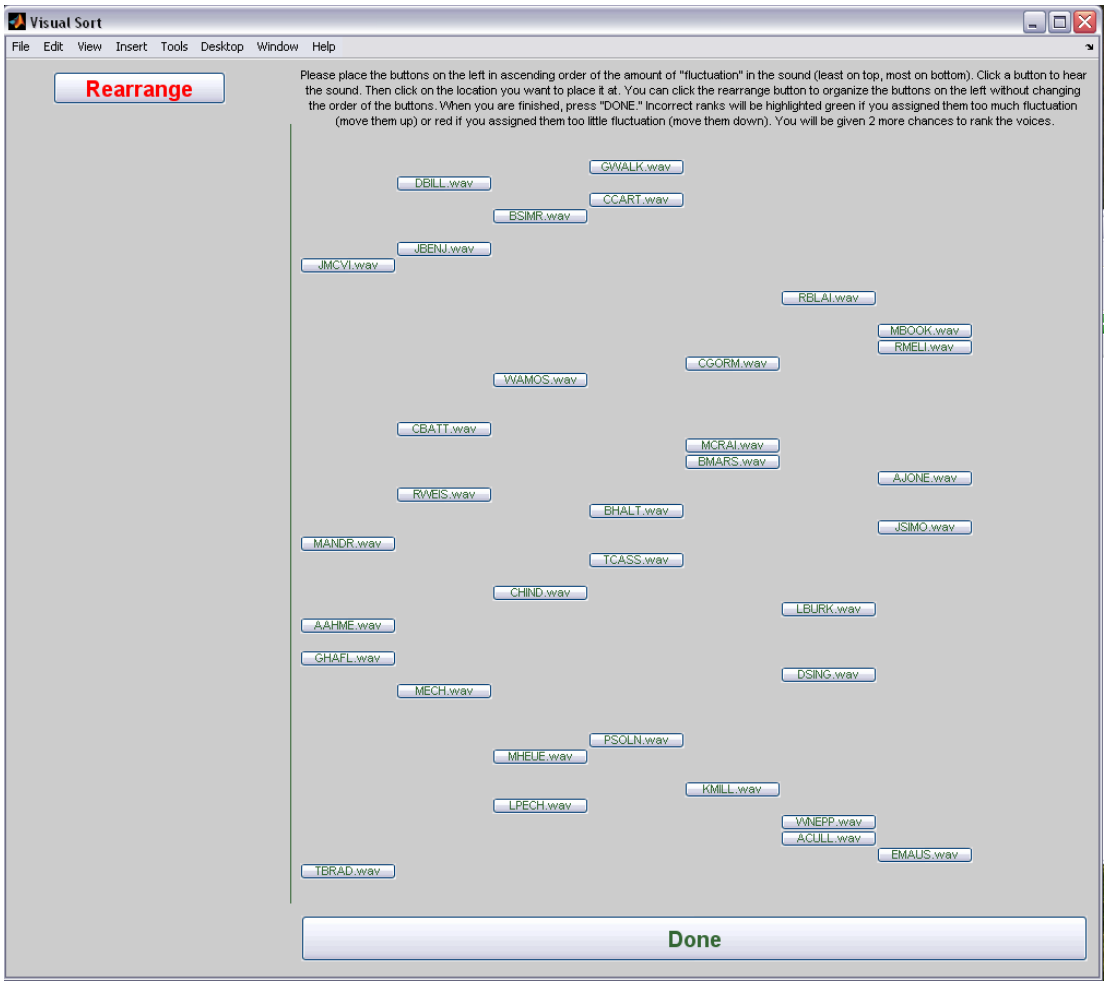

Figure 3-2. The GUI used for the Visual Sort Ranking (VSR) training procedure.

phase listeners evaluated six stimuli which were randomly selected from the Kay Elemetrics Disordered Voice Database (Kay-Elemetrics, Inc., Lincoln, NJ) in order to orient listeners to the sounds used in the experiment. Once the listener demonstrated an understanding of the method by successfully completing the six stimuli set, the adaptive matching task was started. No feedback was provided during this training.

IQD, VSR, and the matching familiarization were all completed in the first session. Subsequent sessions required only that the listener be re-familiarized with the sounds by completing the first section of the VSR training. Two participants were dropped due

to VSR scores which did not meet criteria. Ten listeners completed testing; on average, listeners took three sessions each lasting two hours to complete the experiment.

Method of Constant Stimuli

For the method of constant stimuli, listeners were presented a pair of stimuli and instructed to choose the sound which was ‘more rough.’ The first sound presented was the ‘standard’ to which the second sound, or ‘signal’ was compared. This signal was identical to the reference signal used in the adaptive matching task for breathiness in all respects except for the waveform envelop. When matching breathiness, the SNR of the reference signal was manipulated but the waveform envelop was held constant. In contrast, the reference signal used in the present experiment used a constant SNR, but the modulation depth of a 40-Hz raised cosine amplitude modulation of the waveform envelop was systematically varied. The standard and the signal were represented on a GUI by buttons labeled ‘interval 1’ and ‘interval 2’. The amplitude modulation depth of the signal was varied from -11 dB to -39 dB in steps of 2 dB to represent different levels of roughness. (However, due to a technical error, a modulation depth of -17 dB was not implemented for any of the standards.) Each pair of sounds was presented multiple times in a random order so that each of the standard-signal pairs was evaluated 10 times by each listener.

Training was also conducted for the method of constant stimuli. The IQD and VSR tasks were completed as described previously, at the beginning of each session. Additionally, in the first session, one set of the experimental stimuli were randomly selected to familiarize listeners with the signals and standards used in the experiment. Once again the criteria for participation in the experiment was a score of 0.70 or above on the VSR and a demonstration of understanding the GUI and the experimental

procedures. Once the training was completed the method of constant stimuli task was started. Ten listeners completed the experiment; on average listeners took an average of four sessions each lasting approximately 2 hours to complete the experiment.

CHAPTER 4
RESULTS

Reliability

Inter-judge and intra-judge reliability were calculated for the rating scale and adaptive matching tasks using Pearson's correlation coefficient. Intra-judge reliability for the rating scale task was calculated as the Pearson's correlation between sets of five judgments given by listeners. Intra-judge reliability for the adaptive matching task was calculated as the Pearson's correlation between the thresholds from ascending and descending runs for listeners. Inter-judge reliability for the rating scale and adaptive matching task was calculated by computing the Pearson's correlation coefficient between average rating judgments and average listener thresholds across multiple listeners, respectively.

Table 4-1. Inter- and intra-judge reliability as determined using Pearson's Correlation Coefficient for rating scale and adaptive matching tasks.

	Rating Scale		Adaptive Matching	
	Intra-judge	Inter-judge	Intra-judge	Inter-judge
Mean	0.93	0.91	0.84	0.68
SD	0.04	0.03	0.05	0.05
Range	0.88-0.97	0.88-0.95	0.75-0.90	0.60-0.76

The mean intra-judge and inter-judge reliability for the rating scale task were 0.93 and 0.91 respectively. The mean intra-judge and inter-judge reliability for the adaptive matching task were 0.84 and 0.68 respectively. While these are not necessarily poor scores, these are considerably poorer than those obtained for the breathiness matching data (Patel et al. 2009). The low correlations may be due to several factors.

One possibility is that listeners have difficulty in separating the voice quality of 'roughness' from other percepts such as breathiness or pitch. In order to minimize the

likelihood that listeners were attending to something other than 'roughness', the training paradigm was implemented to specifically demonstrate the differences between possible confounding percepts. However, training does not appear to appreciably increase the intra- and inter-judge reliability for the roughness matching task (Eddins et al., 2009). Another possibility is that judging roughness may simply be harder than judging breathiness, at least in the matching task tested in the present experiment. It is also possible that listeners had difficulty in comparing roughness in voices to that in the reference signal used for the present experiment.

Psychometric Functions for Roughness

The data obtained from the method of constant stimuli are shown in Figure 3. It is evident that this method resulted in a sigmoid function for each standard. As previously described, the method of constant stimuli task typically yields psychometric functions in the form of a sigmoid function. The values obtained generally range from 0 percent correct to 100 percent correct; however, in the present study, the function for each of the standards began to experience a floor effect at approximately the -31 dB modulation depth for the reference signal (Figure 3). Thus, many of the standards never approached the 100 percent correct threshold. One possible explanation for the floor effect which occurred is that the reference signal which was developed performs poorly for stimuli which have very small amounts of roughness. This may be because the reference signal, even at low levels of modulation depth, retains some perceptual characteristics of roughness.

The data for each standard was fit with a logistic function with a customized script written in MATLAB. The logistic function was computed as:

$$P(c) = \frac{1}{1+e^{\frac{\alpha - m}{\beta}}}$$ – Equation 4-1

where m represents the modulation depth (in dB), and α and β were varied until the some of the squared deviations between the predicted and obtained percent correct was minimized. The parameters for each of the standards have been summarized in Table 2.

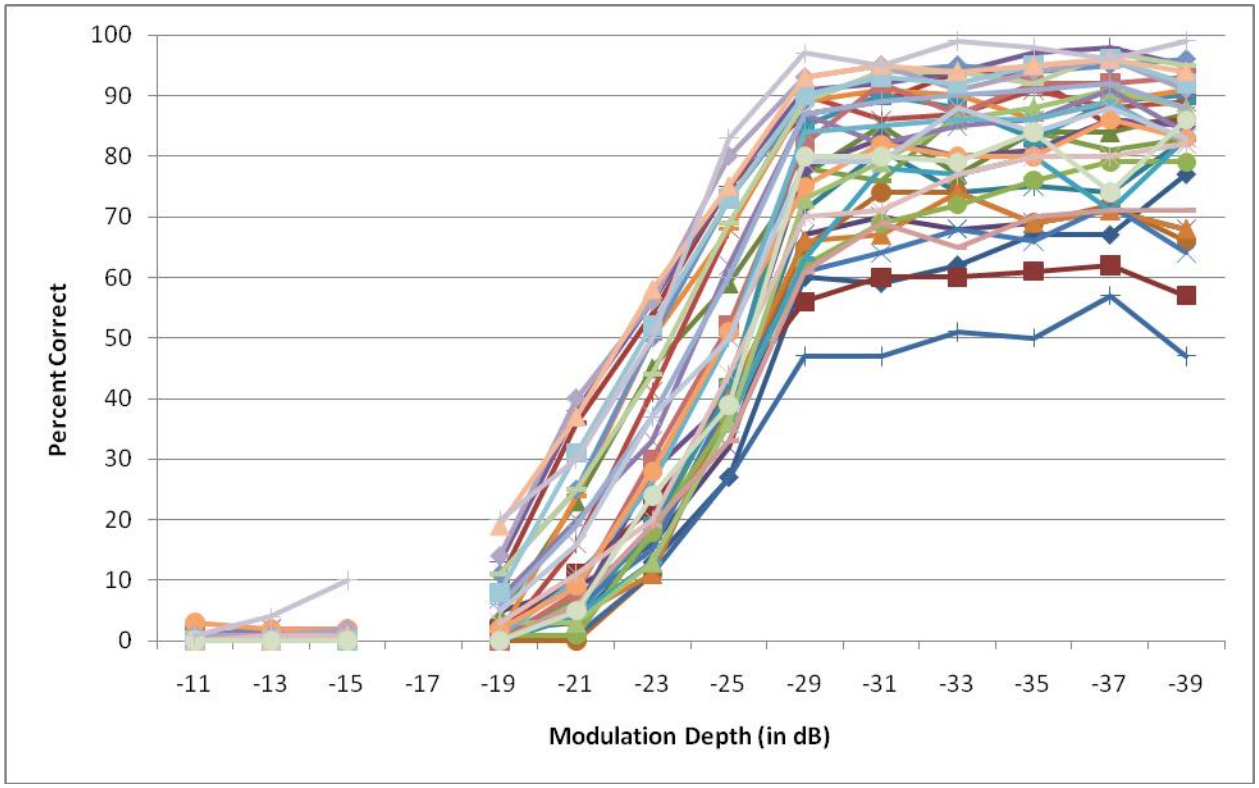

Figure 4-1. Psychometric functions for roughness obtained using the method of constant stimuli. Each line represents one standard. Each point shows the relative number of responses stating that roughness in the signal was greater than roughness in the standard. Data has been averaged across listeners. (Note that data at -17 dB modulation depth was not obtained due to a technical error).

Comparing Perceptual Distances Obtained from the Rating Scale and Method of Constant Stimuli Tasks

In the rating scale task, the average of all ten listeners' ratings was computed for each standard being evaluated. The listener thresholds obtained from the method of

Table 4-2. Parameters for each voice standard calculated using the logistic function described in Equation 1.

Standard	α	β	R^2	Error
AAHME	30.56	3.264018	0.94	0.27
ACULL	31.36	4.027669	0.87	0.35
AJONE	25.15	2.937965	0.96	0.30
BHALT	29.39	3.039287	0.91	0.34
BMARS	27.97	2.457377	0.94	0.34
BSIMR	29.15	2.647865	0.88	0.42
CBATT	34.70	3.866931	0.86	0.33
CCART	22.81	2.353358	0.98	0.26
CGORM	26.87	2.540341	0.96	0.29
CHIND	26.68	2.477156	0.97	0.28
DBILL	26.02	1.883172	0.99	0.24
DSING	29.30	2.968637	0.90	0.39
EMAUS	29.96	3.337962	0.90	0.36
GHAFL	23.92	1.821768	0.99	0.27
GWALK	28.78	3.047885	0.96	0.27
JBENJ	22.53	2.333847	1.00	0.12
JMCVI	28.30	2.66331	0.95	0.31
JSIMO	23.50	2.015154	0.99	0.25
KMILL	23.16	2.101168	1.00	0.12
LBURK	25.30	2.132843	0.99	0.18
LPECH	27.17	2.390649	0.99	0.19
MANDR	24.83	2.521935	0.97	0.27
MBOOK	25.71	2.256526	0.98	0.27
MCRAI	26.43	2.727562	0.96	0.29
MECH	24.46	2.350767	0.99	0.19
MHEUE	29.67	3.129455	0.93	0.31
PSOLN	23.52	2.243659	1.00	0.12
RBLAI	22.29	2.139499	0.99	0.16
RMELI	22.99	2.121762	0.99	0.15
RWEIS	22.36	2.343124	1.00	0.11
TBRAD	25.70	2.99307	0.97	0.24
TCASS	27.63	3.167834	0.96	0.26
WAMOS	26.95	2.402133	0.95	0.33
WNEPP	22.52	2.159864	0.99	0.14
Average	26.40	3.570336	0.96	0.26
SD	3.04	1.412017	0.04	0.08

constant stimuli were also averaged and the modulation depth at which each standard was judged as 'more rough' for 50% of stimuli presentations was calculated from the logistic fit to the data. A correlation of -0.89 was obtained between the average rating scale values and the 50% point on the psychometric curve for that standard. This indicates a strong relationship between the two measures of roughness. This data is shown in Figure 4. A linear regression function was fitted to the data and resulted in an R^2 value of 0.79.

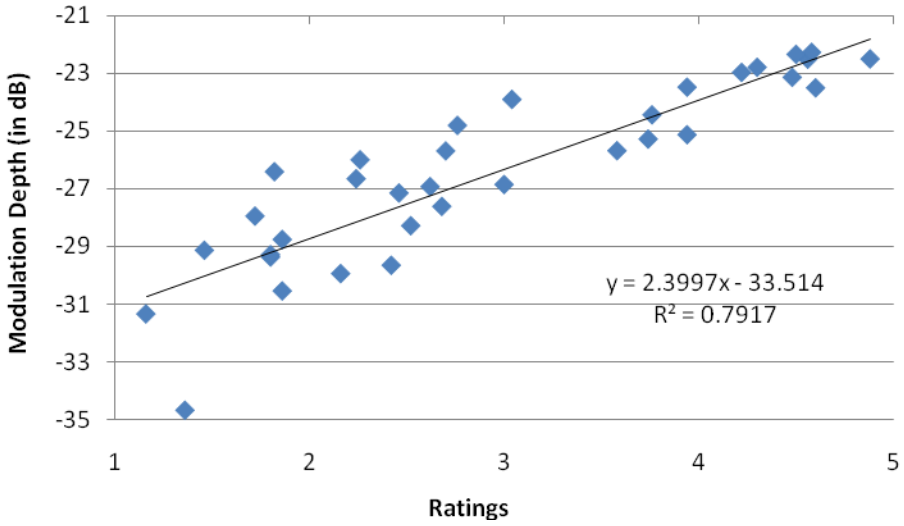

Figure 4-2. Method of constant stimuli 50% thresholds (y-axis) compared to values obtained using the rating scale task (x-axis). The method of constant stimuli thresholds are measured in modulation depth, and the rating scale judgments are measured in listener ratings from 1-5.

Comparing Perceptual Distances Obtained from the Adaptive Matching and Rating Scale Tasks

For the adaptive matching task, the average of the ten listeners' modulation depth thresholds was calculated for each of the 34 standards. A correlation of 0.90 was obtained between the average rating scale values and adaptive matching task listener thresholds. This indicates a strong relationship between these two measures of

roughness as well, and is shown in Figure 5. A linear regression function was fitted to the data, which resulted in an R^2 value of 0.82.

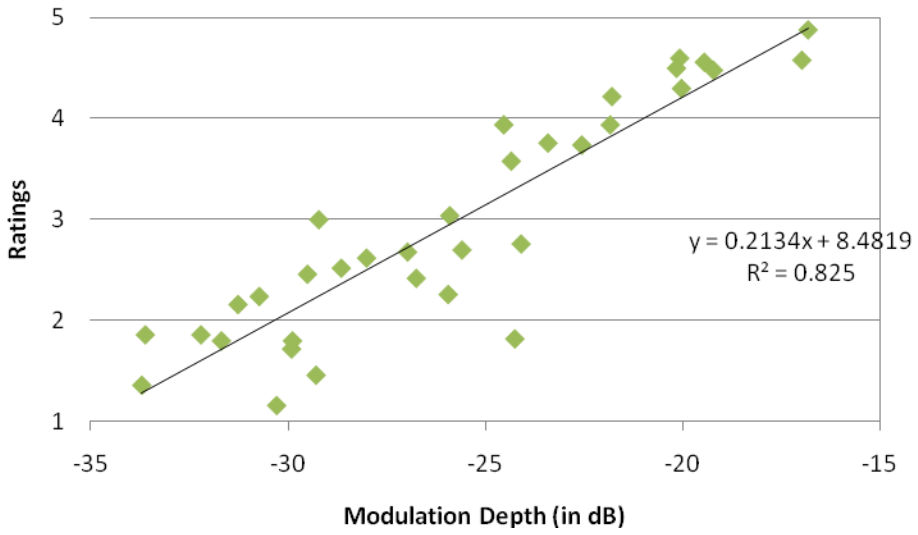

Figure 4-3. Adaptive matching task thresholds (x-axis) compared to the rating scale judgments (y-axis).

Comparing Perceptual Distances Obtained from the Adaptive Matching and Method of Constant Stimuli Tasks

Averaged listener thresholds for the adaptive matching task and 50% thresholds for the method of constant stimuli are also compared, as shown in Figure 6. A correlation of -0.90 was calculated between the adaptive matching task listener thresholds and the 50% point on the psychometric curve for that standard, indicating a strong relationship between the two measures of roughness. A linear regression function was fitted to the data, resulting in an R^2 value of 0.80.

Similar comparisons were also made with other points on the logistic fits to the psychometric functions (30%, 40%, 50%, 60%, 70%, 80% and 90%). These were computed because the method of constant stimuli failed to approximate a 100% score for several standards. These thresholds were then correlated with various other

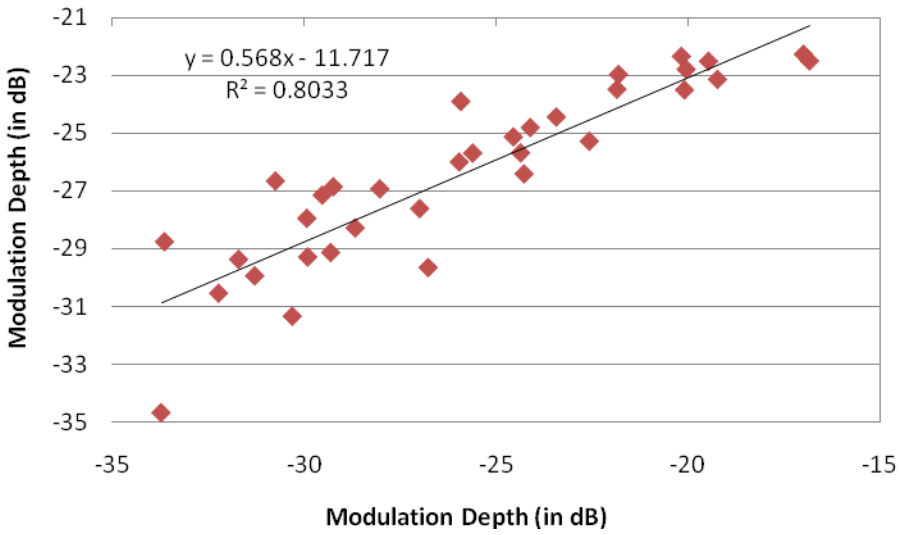

Figure 4-4. Adaptive matching task thresholds (x-axis) versus the method of constant stimuli 50% thresholds (y-axis).

measures of roughness. A summary of the data can be seen in Table 2. Thresholds were correlated with the adaptive matching task thresholds, these ranged from -0.84 to -0.91, with the strongest correlations at the 30% threshold. The 30% threshold also contained the highest correlations for visual sort ranking (VSR) and rating scale comparisons as well, at -0.87 and -0.90 respectively.

Table 4-3. Method of constant stimuli thresholds correlated with values obtained using the adaptive matching task, visual sort ranking, and the rating scale task.

	Adaptive	VSR Rank	Rating
30% Threshold	-0.919	-0.874	-0.902
40% Threshold	-0.910	-0.874	-0.900
50% Threshold	-0.896	-0.865	-0.890
60% Threshold	-0.883	-0.855	-0.879
70% Threshold	-0.869	-0.845	-0.868
80% Threshold	-0.856	-0.835	-0.857
90% Threshold	-0.840	-0.822	-0.843

CHAPTER 5
DISCUSSION

The goal of this study was to evaluate various methods for quantifying rough voice quality in order to validate the use of the adaptive matching procedure for this purpose. A rating scale, adaptive matching task, and the method of constant stimuli were each used by listeners to evaluate the same set of disordered voices. Each method gave similar perceptual data on roughness, as is evidenced by the high correlation scores and R^2 values obtained from comparing each set of data. This suggests that listeners were able to use each method to evaluate the voice standards and produced similar results.

The high level of agreement between methods serves as a basis for suggesting that the adaptive matching task is valid as a measure of rough voice quality. In certain applications it is advantageous to use the adaptive matching task instead of rating scales. This is because the adaptive matching uses a common reference signal across voices and avoids the use of arbitrary numbers to describe the magnitude of a percept. Unlike rating scales, this allows experimenters to compare data across different experiments and listeners. The method of constant stimuli can help provide similar data as well, but it requires greater test time. The additional information obtained through the method of constant stimuli is not always required. However, this approach is useful should an experimenter need to determine psychometric functions for specific stimuli. Unfortunately, the reference signal used in the present experiment failed to give a good match to stimuli with low/mild roughness. This resulted in a saturation of the psychometric functions derived from several stimuli.

While the study did meet the primary goal of validating the adaptive matching task, there are several limitations worth noting. The poor reliability scores observed for the adaptive matching task suggest several possible problems. It may be that perceptual judgments of roughness are simply more difficult for listeners than those for other voice quality percepts. This may be due to the multitude of correlates which are shared with percepts such as breathiness and hoarseness, for example, or to other currently unidentified factors.

One solution was to expose the listeners to stimuli which would span a wide range of rough voices for familiarization. In addition, providing examples of changes in breathiness, pitch and other percepts would help listeners form an internal definition of roughness which excluded these confounds. However, the training paradigm which was implemented for the adaptive matching task (Eddins, et al., 2009) and the method of constant stimuli did not appear to appreciably improve listener performance.

Another possibility is that the “just-noticeable-difference” or the “difference limen” for roughness may be higher than the step sizes used in the present experiment, at least for voices with low levels of roughness. This would result in greater variability in the adaptive matching thresholds for voices with low levels of roughness. Moreover, the reference signal which was designed for this experiment may need to be refined for evaluating voices containing less roughness. A few other reference signals were attempted in pilot work, but were discarded as they did not provide a wide enough range of roughness (Shrivastav, 2008). These included (i) a sine wave carrier with a sine-wave amplitude modulation, (ii) a sawtooth+noise carrier with a sine wave amplitude modulation, (iii) and a sawtooth+noise carrier with a square-wave amplitude modulation.

Each of these possibilities would require further research before they could be implemented in the current work on measuring roughness.

Although there are several limitations which hinder the effectiveness of this method, the average adaptive matching threshold across listeners appears to be stable. If the stability of the average matching threshold can be further confirmed, the adaptive matching task can still have great utility in applications such as the development of voice quality perception models. The data produced is easier to translate across experiments, making it useful for models which utilize average data.

The logistic functions derived for the 34 voices show considerable variability in their parameters. One interpretation of this finding is that the perception of roughness varies across voices – perhaps affected by factors such as pitch or other perceptual attributes of the voice. However, the variability in these parameters more likely results from the failure of several voices to approach 100% scores when matched to the reference signal used in this study. This introduced some error in the resulting logistic functions, making a direct comparison of fit parameters difficult. To address this limitation, some preliminary analyses were undertaken. A new logistic fit was derived for each voice stimulus, but only using data until -31 dB modulation depth of the reference signal. This was done to remove data points where the listener judgments saturated against the reference (i.e. the roughness of the reference signal did not change with further decrease in modulation depth). As expected, the variability in α and β was lower for these functions than for those obtained using all data points. However, it remains difficult to determine whether the variability in these parameters reflects true

differences in psychometric functions or are an artifact of the floor effect obtained with the reference signal used here. Further experimentation is necessary to answer this question.

CHAPTER 6
CONCLUSIONS

The present experiment compared three methods for measuring the voice quality roughness. The rating scale, adaptive matching task, and method of constant stimuli were each used to evaluate the same 34 stimuli with varying degrees of roughness. In each of the tasks listeners were able to comparably indicate the amounts of roughness contained in the stimuli. Each of the methods gave similar perceptual distances between stimuli; however, the adaptive matching task remains the ideal method for quantifying rough voice quality for several reasons. Firstly, the task removes many of the biases and variability which are associated with perceptual data collection, and found specifically in rating scales. Secondly, the method, when completed with multiple judgments per stimulus, yields data which correlates well in inter- and intra-judge data. Thirdly, in comparison with other methods which meet the above criteria, the adaptive matching task is more efficient, providing comparable accuracy without the additional concerns associated with listener fatigue, limited resources, and time constraints. Lastly, the data yielded by the adaptive matching task offers good resolution and a unit with which to measure roughness (modulation depth measured in dB or percent).

LIST OF REFERENCES

Dai, H. (1995). On measuring psychometric functions: a comparison of the constant-stimulus and adaptive up-down methods. *Journal of the Acoustical Society of America, 98(6),* 3135-3139.

Deal, R., & Emanuel, F.W. (1978). Some waveform and spectral features of vowel roughness. *Journal of Speech, Language, and Hearing Research, 21,* 250-264.

Eadie, T. L., & Doyle, P. C. (2002). Direct magnitude estimation and interval scaling of pleasantness and severity in dysphonic and normal speakers. *Journal of the Acoustical Society of America,112(6),* 3014-3021.

Eddins, D.A., Shrivastav, R., Patel, S. (2009). "Does training improve consistency of roughness judgments in a matching task?" Presented at the 157th Meeting of the Acoustical Society of America, May 18-22,2009, Portland, OR.

Eskanazi, L., Childers, D.G., & Hicks, D.M. (1990). Acoustic correlates of vocal quality. *Journal of Speech and Hearing Research, 33,* 298-306.

Heiberger, V.L., & Horii, Y. (1982). Jitter and shimmer in sustained phonation. In N.J. Lass (Ed.), Speech and language: Advances in basic research and practice (Vol. 7, pp. 299-332). New York: Academic.

Hirano, M. (1981). Clinical examination of voice. Wien ; New York: Springer-Verlag.

Hirano, M., Hibi, S., Yoshida, T., Hirade, Y., Kasuya, H., & Kikuchi, Y. (1988). Acoustic analysis of pathological voice*. Acta Otolaryngologica (Stockholm), 105,* 432-438.

Kreiman, J., Gerratt, B. R., Kempster, G. B., Erman, A., & Berke, G. S. (1993). Perceptual evaluation of voice quality: review, tutorial, and a framework for future research. *Journal of Speech and Hearing Research, 36(1),* 21-40.

McCaffery, M., & Pasero, C. (1999). Pain : clinical manual (2nd ed.). St. Louis: Mosby.

Moore, B.C.J., Glasberg, B.R., & Baer, T. (1997). A Model for the Prediction of Thresholds, Loudness, and Partial Loudness. *Journal of the Audio Engineering Society, 45(4),* 224-240.

Patel, S., Shrivastav, R., & Eddins, D. A. (2009). Perceptual Distances of Breathy Voice Quality: A Comparison of Psychophysical Methods. *Journal of Voice, 24(2)*, 168-177.

Shrivastav, R., Sapienza, C. M., & Nandur, V. (2005). Application of psychometric theory to the measurement of voice quality using rating scales. *Journal of Speech, Language, and Hearing Research*, *48(2),* 323-335.

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

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

Stevens, S.S. (1972). Perceived Level of Noise by Mark VII and Decibels (E). *Journal of the Acoustical Society of America*, *51(2B),* 575-601.

Titze, I. R. (1994). Principles of Voice Production (pp. 354). Englewood Cliffs, NJ: Prentice Hall.

Zwicker, E., & Fastl, H. (1990). Psychoacoustics, facts, and models. Berlin ; New York: Springer-Verlag.

BIOGRAPHICAL SKETCH

Stacie Cummings is a graduate of the master's program focusing on Speech-Language Pathology at the University of Florida Department of Communication Sciences and Disorders. During her master's program, she completed a master's thesis on the comparison of psychophysical methods for the evaluation of rough voice quality under the mentorship of Rahul Shrivastav, Ph.D. The thesis was accepted for a poster presentation at the 39th Annual Symposium of the Voice Foundation meeting in 2010. Ms. Cummings is a previous graduate of the University of Florida, earning her B.A. in Linguistics with a minor in the Teaching of English as a Second Language in December 2007. She was awarded the University Scholars research scholarship in the summer of 2007, and the resulting research on breathy voice quality was accepted as a poster presentation at the 37th Annual Symposium of the Voice Foundation in 2008. Ms. Cummings worked as a research assistant in the Voice Acoustics and Perception Laboratory, under the supervision of Rahul Shrivastav Ph.D. Her duties included participant recruitment, data collection and analysis, literature review, and preparation for research presentations. She presented research on rough voice quality in a poster session at the 156th meeting of the Acoustical Society of America in 2009. In the summer of 2010 Ms. Cummings plans to begin her clinical fellowship year.