The influence of motor production experience on voice perception

A. Louise Pinkerton

University of Iowa
THE INFLUENCE OF MOTOR PRODUCTION EXPERIENCE ON VOICE PERCEPTION

by

A. Louise Pinkerton

A thesis submitted in partial fulfillment of the requirements for the Master of Arts degree in Speech Pathology and Audiology in the Graduate College of The University of Iowa

August 2016

Thesis Supervisor:  Associate Professor Eileen Finnegan
CERTIFICATE OF APPROVAL

MASTER'S THESIS

This is to certify that the Master's thesis of

A. Louise Pinkerton

has been approved by the Examining Committee for the thesis requirement for the Master of Arts degree in Speech Pathology and Audiology at the August 2016 graduation.

Thesis Committee:

Eileen Finnegan, Thesis Supervisor

Ingo Titze

Meredith Saletta
To my family for their love and support.
Acknowledgements

My sincere thanks and appreciation to my thesis advisor and mentor Dr. Eileen Finnegan for her guidance and support throughout my degree and the thesis process.

My grateful thanks to Dr. Ingo Titze and Dr. Meredith Saletta who served on my committee and provided input and insight on the experiment design and this document.

Thank you to Dr. Michael Karnell for the foresight to collect voice samples at the UIHC Voice Clinic and the willingness to share the database with me for my thesis.

Additionally, I would like to thank the many professors at the University of Iowa Department of Communication Sciences and Disorders who provided an exceptional educational experience, especially Anne Wallace and Ann Fennell, and the professors at Indiana University who first introduced me to voice science and vocal pedagogy, Dr. Moya Andrews, Roy Samuelson, and Paul Kiesgen.

My many thanks to the staff with the Department of Communication Sciences and Disorders for their assistance throughout my degree and the thesis process.

Thank you to my friends and classmates that supported me throughout our studies, especially to Shannon Ho, our Iowa vocology cohort, my clinical team members, and the other thesis students.

Thank you to Dr. Yang Zhang for his guidance on statistics and to Kristin Gibson for editing.

And finally, thank you to all the students and experts who took time from their busy lives to participate as subjects. This study would not have been possible without you.
Perceptual speech and voice analysis is an essential skill for all speech-language pathologists, but it is a difficult skill to teach. Even the reliability for experienced experts is variable. Some training literature and practices in speech-language pathology suggest that imitating pathological voices may be useful for developing perceptual judgment. Evidence from other fields suggests that motor experience influences perception. Until now the link between production and perception of voice quality has not been addressed. The purpose of this pilot study is to test the hypothesis that imitating pathological voice samples would improve the perceptual discrimination abilities of naïve, inexperienced listeners.

Three expert listeners rated 25 voice samples using a perceptual voice evaluation scale, the Grade, Instability, Roughness, Breathiness, Asthenia, Strain Scale (GIRBAS) (Dejonckere et al., 1996), and identified anchor samples for the training protocol. These expert ratings were used to develop summary expert ratings that served as a comparison for the naïve listener ratings. Two groups of naïve undergraduate listeners received training in evaluating voice quality and in administering the GIRBAS. They completed a pretest, a training session, a homework session, and a post-test. During each activity, they rated 6 voices and provided a confidence rating for their scores. The experimental group imitated the voice samples during the study, and the control group completed the training without supplemental motor experience.

It was hypothesized that both listener groups would have improved accuracy and confidence levels between the pretest and post-test, with a larger improvement for the experimental group. Data suggested that training improved naïve listener accuracy and confidence levels and that this improvement was maintained for at least seven days after the initial training. Post-test accuracy for both groups was approximately the same. Imitation did not improve the accuracy of ratings, although those subjects had higher confidence levels. The data supported previous research that found that training improved the accuracy of perceptual voice evaluations. However, the hypothesis that imitation could improve perceptual ratings was not supported by this study and bears further investigation due to the small sample size.
Public Abstract

As part of their clinical work diagnosing and treating people with voice disorders, speech-language pathologists listen to and judge voice quality. This is a challenging skill to teach because these judgments are subjective, and even experienced experts do not exactly agree when rating the same voices. Research indicated that listener training improved agreement, but research has not looked at whether imitating voices with disorders would help students improve how they listen to and judge voice quality. This idea is endorsed by clinicians and professors, but has not been researched.

A group of three experienced speech-language pathologists rated voice samples, and those ratings were used to create a summary expert rating for each sample. Two groups of naïve listeners, a control group and an experimental group, rated voices in a pretest, a training session, a homework assignment, and a post-test session. The control group received basic training alone, and the experimental group received basic training and imitated the voice samples. This study evaluated whether training improved the accuracy and confidence levels of the student listeners when compared to the summary expert ratings and whether imitation improved accuracy and confidence levels when compared to the control group.

Results showed that training did improve the accuracy of student listener ratings and confidence levels for both groups. In contrast to our original idea that the imitation group would be more accurate, the accuracy of the imitation and control groups was approximately the same on the post-test. Both groups were more confident on the post-test, but the confidence level of the imitation group increased the most. But, this was a small study and a study with a larger group of participants would have more conclusive results.
# Contents

List of Tables .......................................................................................................................... vii
List of Figures .......................................................................................................................... viii

The Influence of Motor Production Experience on Voice Perception ........................................ 1

Literature Review ..................................................................................................................... 4
  Overview of Graduate Student Preparation in Voice ............................................................... 4
  Perceptual Voice Evaluations ................................................................................................. 6

The Influence of Motor Learning on Voice Perception ............................................................ 11

Theoretical Models and Neurological Research ...................................................................... 13

Methodology .......................................................................................................................... 19
  Study Participants ................................................................................................................ 19
  Expert Listener Sessions, Anchor Samples and Summary Ratings ........................................ 20
  Naïve Listener Sessions ....................................................................................................... 23
  Data Analysis ....................................................................................................................... 25

Results ................................................................................................................................... 28
  Participant Group Demographics ......................................................................................... 28
  Expert Results ...................................................................................................................... 30
  Naïve Listener Results ........................................................................................................ 32

Discussion ............................................................................................................................. 42
  Influence of Training ........................................................................................................... 42
  Influence of Imitation .......................................................................................................... 43
  Naïve Listeners’ Narrative Responses .................................................................................. 46

Limitations ............................................................................................................................... 48

Future Directions ................................................................................................................... 49

Conclusions ............................................................................................................................. 51

References ............................................................................................................................... 52
List of Tables

Table 1 Expert Listener Demographics ................................................................. 28
Table 2 Naïve Listener Demographics .................................................................. 29
Table 3 Naïve Listener Demographics by Experimental Group ............................. 30
Table 4 Expert Listener Agreement for Voice Samples ......................................... 31
Table 5 Expert Listener Agreement for Anchor Samples ....................................... 31
Table 6 Naïve Listener Confidence Levels Pre- and Post-test ............................... 38
Table 7 Naïve Listener Post-test Confidence Levels ............................................ 40
Table 8 Naïve Listener Narrative Responses ...................................................... 41
List of Figures

Figure 1 A model of factors influencing voice quality ratings. .................................................................8
Figure 2 Dual-stream model of speech perception. .....................................................................................15
Figure 3 The linked dual representation model.........................................................................................16
Figure 4 Naïve listener training sessions. ..................................................................................................25
Figure 5 Naïve listener divergence pretest to post-test by voice quality. .................................................33
Figure 6 Naïve listener divergence by voice quality and session. ..............................................................34
Figure 7 Naïve listener average divergence by subject and session. .........................................................35
Figure 8 Naïve listener divergence pretest to post-test by voice quality and subject. .............................37
Figure 9 Naïve listener post-test divergence by voice quality and experimental group. ..........................39
The Influence of Motor Production Experience on Voice Perception

Perceptual speech and voice analysis is an essential skill for all speech-language pathologists, but it is a difficult skill to teach. Persons with voice disorders are a low incidence population, and consequently, graduate students have limited training in—and exposure to—voice disorders. This makes it challenging for students to become proficient and reliable in evaluating voices. Even the reliability of judgments made by experienced raters is often variable; however, perceptual training programs for experienced and inexperienced raters have reduced intra- and inter-rater variability (Chan & Yiu, 2002; Eadie & Baylor, 2006; Ghio et al., 2014; Yamaguchi et al., 2003). Some training literature and practices in speech-language pathology suggested that imitation of disordered speech or voice quality can be useful for developing perceptual judgment skills (DeBoer & Shealy, 1995; Duffy, 2013), but this premise has not been evaluated by the research literature on perceptual voice evaluations. Evidence in other fields, such as motor speech production and neurology, suggested that motor experience influences perception (Aglioti & Pazzaglia, 2010; Baese-Berk & Goldrick, 2009; Chang, Kenney, Loucks, Poletto, & Ludlow, 2009; Kleber, Veit, Birbaumer, Gruzelier, & Lotze, 2010; Lévêque, Muggleton, Stewart, & Schön, 2013; Zarate & Zatorre, 2008), and several theories of speech production make a connection between motor production and comprehension of speech (Galantucci, Fowler, & Turvey, 2006; Hickok & Poeppel, 2000; Hutchins & Moreno, 2013; Liberman & Mattingly, 1985). Until now the link between production and perception had not been investigated in the area of voice quality. The purpose of this pilot study was to test the hypothesis that imitating pathological voice samples would improve the perceptual discrimination abilities of naïve listeners.

Subjective perceptual evaluations are a staple of clinical voice assessment procedures and are used widely for describing the quality of the patient’s voice and quantifying the severity of the disorder (Ghio et al., 2014). Standard perceptual assessments, like the Consensus Auditory-Perceptual Evaluation – Voice (CAPE-V) (Kempster, Gerratt, Verdolini Abbott, Barkmeier-Kraemer, & Hillman, 2009), the Grade, Roughness, Breathiness, Asthenia, Strain (GRBAS) scale (Hirano, 1981) and its derivative the Grade, Instability, Roughness, Breathiness, Asthenia, Strain (GIRBAS) scale (Dejonckere et al., 1996),
use general descriptors of voice quality. Raters evaluate the voice on different voice qualities and indicate the severity for each quality on a 100 mm line (CAPE-V) or on an ordinal scale from 0–3 (GRBAS and GIRBAS). These descriptive terms seem to have face validity but are not supported by a theory of voice production (Kreiman, Gerratt, Kempster, Erman, & Berke, 1993). Furthermore, perceptual voice quality rating scales depend on a clinician’s experience and subjective judgment to describe a voice sample. Research found inconsistent intra- and inter-rater reliability in subject ratings using these scales, and these scales correlated only weakly with objective instrumental measures (Kreiman et al., 1993). Objective measures may support perceptual evaluations, but the ear remains the gold standard for evaluating describing the voice. Therefore, it is important to consider the most effective methods for training speech-language pathologists (SLPs) to make these subjective judgments.

This study investigated the influence of motor training on perceptual skills in naïve listeners to consider additional methods to effectively train Master’s SLP students in the clinical judgment of voice quality. The preparation of Master’s SLP students to evaluate and treat voice varies widely (Teten, DeVeney, & Friehe, 2013; Van Mersbergen, Ostrem, & Titze, 2001) and new clinicians did not feel confident in their abilities (Teten, DeVeney, & Friehe, 2016). Research studies on vocal perceptual evaluations looked further into why perceptual ratings were less reliable and what could be done to improve them. These studies found that reliability was better for the least and most severe voice samples (Ghio et al., 2014; Kreiman, Gerratt, & Ito, 2007), listener experience influenced reliability (De Bodt, 1997; Ghio et al., 2014; Karnell et al., 2007; Kreiman, Gerratt, & Precoda, 1990; Sofranko & Prosek, 2014) and training programs for naïve listeners were effective in improving reliability (Awan & Lawson, 2009; Chan & Yiu, 2006; Eadie & Baylor, 2006; Ghio et al., 2014; Iwarsson & Reinholt Petersen, 2012; Yamaguchi, Shrivastav, Andrews, & Niimi, 2003).

Based on models of production and perception in speech and voice, as well as research on mirror neurons, it was proposed that motor experience developed by imitating pathological vocal quality would influence a listener’s skills in discriminating between pathological voice qualities, thereby enhancing training. The motor theory of speech perception (Liberman & Mattingly, 1985) and the dual-stream
models of speech and voice perception (Hickok & Poeppel, 2000; Hutchins & Moreno, 2013) support the idea that motor knowledge influences our perception of the voice, and therefore, our perception of voice quality. The hypothesis about the role of motor experience was reinforced by studies of pitch perception (Hutchins & Moreno, 2013), neural activation when listening to singing (Lévêque et al., 2013) and vocal motor training (DeBoer & Shealy, 1995). However, only one study of perceptual voice evaluations has included motor experience (Iwarsson & Reinholt Petersen, 2012), and none have evaluated the role of motor experience on improving raters’ accuracy. The role of motor experience in stabilizing or training the listener’s internal standards for voice quality has not been specifically researched.

**Research question.** The question addressed in this study was whether experience imitating pathological voices would influence a listener’s perceptual discrimination skills and confidence level when evaluating the same pathological voice qualities. Two groups of naïve listeners completed a pretest on discriminating voice quality and were trained in categorizing voice quality with the GIRBAS. The first group received traditional didactic training and voice rating practice, and the second group received that same training, plus experience imitating pathological voice qualities. Naïve listener ratings were compared to summary GIRBAS ratings developed from the individual expert ratings. This research provided preliminary information about whether vocal motor experience (imitation) influenced a listener’s perceptual evaluation of other voices. More specifically:

1. Does training improve accuracy of ratings and subject confidence level (i.e., are post-test scores higher than the pretest scores for naïve listeners)?

2. Does imitation improve accuracy of ratings and subject confidence level in comparison to the control (no imitation) condition?
Literature Review

The literature review presents information about Master’s SLP student preparation in voice, inconsistencies in vocal perceptual evaluations, research into motor learning and voice perception, and theoretical models of perception and production with neurological support for those models. A review of graduate student preparation identifies the need to investigate effective training methods, such as imitation, to efficiently prepare students for their careers. Prior research on vocal perceptual evaluations points to reasons for inconsistent reliability and identifies protocols to improve reliability that were applied in this study. The literature review presents details of two studies that examined the influence of imitation or motor learning on vocal perceptual skills of SLP students. These studies identify what we already know about imitation and perception and clarify how this study advances this topic. The final section introduces theoretical models and neurological data supporting the premise that imitation experience can mediate perception, which is the fundamental concept evaluated in this study.

Overview of Graduate Student Preparation in Voice

Speech-language pathologists will likely encounter clients with voice disorders in their careers and may not have the skills necessary to evaluate and treat these clients. Voice disorders are low incidence disorders, yet voice clients are common in multiple settings—from the schools to clinics and hospitals. The prevalence of voice disorders among adults throughout their lifetime is 29.9% and the incidence at a given time is 6.6% (Roy, Merrill, Gray, & Smith, 2005). Voice disorders occur in 14% of the pediatric population (Bhattacharyya, 2015). In health care settings, 5% of the clinical load is adult voice clients, and 2% is pediatric clients (American Speech-Language-Hearing Association, 2015). In the public schools, 22% of surveyed clinicians regularly provided voice treatment to an average of 1.5 clients per year (American Speech-Language-Hearing Association, 2014).

Coursework. For Master’s SLP students, the proportion of coursework and clinical experiences dedicated to voice habilitation and voice disorders varies widely. Students may be briefly introduced to the study of voice disorders at the undergraduate level in general coursework on communication sciences and disorders, but the bulk of these topics are completed at the Master’s level. The American Speech-
Language-Hearing Association (ASHA) requires applicants for the Certificate of Clinical Competency (CCC-SLP) to “have demonstrated knowledge of communication and swallowing disorders and differences, including the appropriate etiologies, characteristics, anatomical/physiological, acoustic, psychological, developmental, and linguistic and cultural correlates” in the area of voice and resonance (Council for Clinical Certification in Audiology and Speech-Language Pathology of the American Speech-Language-Hearing Association, 2013). In 71% of graduate programs, this knowledge is acquired through only 1–3 credits of coursework on voice disorders (Van Mersbergen et al., 2001). Some graduate programs offer additional coursework in voice habilitation, voice production, and professional voice, but these programs are rare.

Because Master’s students may have limited coursework in voice and voice requires a unique knowledge set, it is important to be efficient in training students to evaluate voice quality. Master’s students do take courses in other specializations that use speech and voice evaluations, such as motor speech disorders, articulation and fluency. However, the vocabulary and evaluation criteria for perceptual evaluations varies significantly even from voice to motor speech disorders, which are closely related. The use of imitation could enhance the training students do receive.

**Clinical experience.** Additionally, Master’s students may have limited clinical exposure to clients with voice disorders during their studies. Because this is a low incidence population, it can be challenging for academic clinics to recruit and retain enough voice clients and to provide appropriate supervision to offer all students experiences treating this population (Teten et al., 2013). Direct clinical experience with voice clients is not required for the Certificate of Clinical Competency through ASHA (Council for Clinical Certification in Audiology and Speech-Language Pathology of the American Speech-Language-Hearing Association, 2013). Graduate programs varied from requiring no clinical hours in voice (27% of programs) to over 25 hours (1%) (Van Mersbergen et al., 2001). Voice is a specialized field, and there is little crossover in the specific techniques and methods for evaluating and treating voice clients. General clinical skills learned through clinical experiences with other populations are helpful, but are not a sufficient clinical foundation for treating voice.
**Career preparation.** Voice evaluation and treatment have been identified as areas in which clinicians and their employers feel they are underprepared. School and medical SLPs in Nebraska indicated that they did not feel prepared to evaluate and treat voice disorders (Teten et al., 2016). The medical SLPs’ confidence level ($M = 2.41$, $1 = $minimally prepared$ and $5 = $extremely prepared$) was higher than the school SLPs’ confidence level ($M = 1.94$). But neither group averaged above “moderately prepared.” The clinicians who feel most confident treating voice were those who completed continuing education in voice, had recently treated voice clients, and/or had a higher percentage of voice clients, plus those who already felt more prepared when they initially graduated. These findings are consistent with a New Zealand survey where both new clinicians and their employers identified voice therapy as an area of weakness (Tillard, Lawson, & Emmerson, 2011).

In summary, it is important to investigate methods to efficiently prepare SLPs for evaluating and treating voice disorders because the typical SLP graduate will complete one course in voice disorders, have limited or no clinical experience in voice, and feel less than moderately prepared in this specialized area. Improving training techniques for perceptual evaluations with techniques such as imitation is one possible step in that process.

**Perceptual Voice Evaluations**

The following section reviews research on rater reliability in expert listeners, identifies reasons for the variability and discusses methods for improving rater reliability. The research on reliability indicates that inter- and intra-rater reliability remains variable, and there is value to considering new training protocols such as imitation. An examination of reasons that raters are not consistent provides insight into perceptual evaluations and how imitation might help stabilize the internal representations of voice qualities, therefore improving reliability. Finally, research on anchor samples and training methods that have been effective in improving reliability and were used in this study are reviewed.

**Inconsistencies in perceptual voice evaluations.** Research into the reliability of perceptual voice evaluations continued to demonstrate inconsistent results in inter- and intra-rater reliability of judgments made with perceptual rating instruments, like the Grade, Roughness, Breathiness, Asthenia,
Strain (GRBAS) scale. For example, recent studies found fair to good reliability for grade, roughness, breathiness, asthenia, and strain (Dejonckere et al., 1996), acceptably high reliability for the entire instrument (Webb, 2004), acceptable agreement for grade (Karnell et al., 2007), and high reliability for the GRBAS and CAPE-V (Nemr et al., 2012). In contrast, other findings included low reliability for voice qualities evaluated by the GRBAS protocol: low agreement for instability (Dejonckere et al., 1996), only moderate reliability with the exception of grade (De Bodt, Wuyts, Van de Heyning and Croux, 1997), and low agreement for asthenia and strain (Yamaguchi et al., 2003). Dejonckere et al. (1996) analyzed voice ratings at five different institutions and found that intra-rater reliability was higher than interrater reliability. Sellars et al. (2009) found only one of three experienced clinicians made consistent ratings, and raised concerns about the widespread adoption of the GRBAS without further investigations into methods for increasing reliability. Studies which tested methods to improve reliability of the GRBAS, to be discussed below, also found inconsistent reliability for various voice qualities.

The reliability varied depending on the voice quality evaluated, protocol variations, training procedures, and the rating instrument. Protocols for the studies with higher interrater reliability included training components or extensive training, focused on qualities known to have higher rater reliability, and correlated the GRBASs with another subjective voice evaluation. The studies finding low interrater reliability had less training, high criteria for reliability, and evaluated qualities known to have low reliability, such as asthenia. Despite these limitations, the GRBAS remains one of the standard vocal perceptual evaluation tools used internationally in clinical and research settings (Yamaguchi et al., 2003).

**Reasons for continued variability.** In general, the validity of perceptual evaluations of voice quality was challenged because they were not based on theoretical models of voice production (Kreiman & Gerratt, 2010; Kreiman et al., 2007; Kreiman et al., 1993). Kreiman et al. (1993) proposed a model of factors that influence the evaluation of an acoustic signal with a perceptual rating scale (Figure 1).
The proposed model considers the general quality of the voice signal, listener factors, task factors, and interaction effects between the factors as the listener rates voice quality. Four of these factors together accounted for 84.2% of the variance in listener agreement: unstable internal listener references, difficulty isolating the voice quality being rated, resolution of the rating scale and assigning mid-level severity ratings (Kreiman et al., 2007). Further discussion follows on the influence of listener factors and task factors on reliability in relation to the training protocol developed for this study.

Listener factors. Listener factors influencing reliability included unstable internal representations of voice qualities, the listener’s professional field (e.g., singing teacher or SLP), listener bias, sensitivity to aural stimulus, and errors. Kreiman, Garrett and Ito (2007) identified unstable criteria for severity levels as one factor that produced listener disagreement. Ghio et al. also found that “unstable and listener-specific prototypes” accounted for disagreements between listeners (2014, p. 2). Additionally, professional background, SLP versus singing teacher, was more influential than the number of years in the profession, with singing teachers rating voices as more severely dysphonic than the SLPs (Sofranko & Prosek, 2014). Listener self-perception created bias in perceptual ratings because listeners often used their own voice as a baseline (Ghio et al., 2014; Kreiman et al., 1990; Kreiman, Gerratt,
Precoda, & Berke, 1992) and did not objectively evaluate their own voice in comparison to others (Haskell, 1987).

A listener’s ability to discriminate and perceive subtle differences in voice quality is not known and may be related to difficulties in rating the severity of voice qualities, such as grade, roughness, etc. Inexperienced listeners were successfully trained to identify isolated 5 dB level changes in the aspiration level in synthesized breathy voice samples (Chan & Yiu, 2006). However, the midpoint ratings on the GRBAS were most challenging for listeners to differentiate (Ghio et al., 2014; Kreiman et al., 2007). Listener error also affected reliability of perceptual ratings by introducing uncontrolled random errors including listener fatigue, notational and protocol errors, lack of attention and the like (Kreiman et al., 1993; Shrivastav, Sapienza, & Nandur, 2005).

**Task factors.** Task factors influencing reliability included the rating scale, quality definitions, listening context, the perceptual task and the speech sample. Listener agreement varied depending on the type and resolution of rating scale used for perceptual evaluations. Reliability was better for an ordinal scale, modified visual analog scale, and a continuous scale with reference samples (Kreiman & Gerratt, 2010; Nagle, Helou, Solomon, & Eadie, 2014; Webb et al., 2004; Yu, Revis, Wuyts, Zanaret, & Giovanni, 2002). The voice has many different qualities and can be considered a multidimensional sound signal, and aspects of the signal interact during perceptual evaluation. It was found that breathiness influenced the rating of roughness (Kreiman, Gerratt, & Berke, 1994), and listeners had difficulties isolating different qualities in a natural voice sample versus synthetic (Gerratt, Kreiman, Antonanzas-Barroso, & Berke, 1993; Kreiman et al., 2007). Furthermore, voice quality terms are not universally defined and do not generally correlate with objective acoustic measures, which confounds the rating task (Awan, Roy, Jette, Meltzner, & Hillman, 2010; Dejonckere et al., 1996; Eadie & Baylor, 2006; Kreiman & Gerratt, 2010; Kreiman et al., 1994; Yu et al., 2002).

The rating task, presentation of samples, and type of speech sample are additional factors producing variability in perceptual ratings. The type of perceptual comparison task—for example, paired-comparison, reference matching or analysis-by-synthesis—influenced the reliability of perceptual voice
ratings (Chan & Yiu, 2002; Gerratt et al., 1993; Kreiman et al., 2007). Kreiman et al. (1993) observed that listener ratings were influenced by the severity of surrounding voice samples and that ratings “drifted” between the first and second ratings of a sample during a single rating session (Gerratt et al., 1993). Lu and Matteson (2014) investigated the influence of different speech samples on the agreement of grade, roughness and breathiness ratings. They found that counting from 1 to 10 and a sustained /a/ at different pitch levels had the highest interrater reliability.

**Methods to improve reliability.** Experimental studies identified methods to improve reliability related to listener factors and task factors that produce variability in ratings. Some research that identified methods related to task factors to improve reliability were reviewed in the previous section. Findings about anchor samples and listener training applicable to this study follow.

**Reference (anchor) samples.** Reference-matching tasks with anchor samples improved reliability. The improvement may be related to providing consistent external references, isolating the features to be evaluated, and reducing the effect of sample presentation. Gerratt et al. (1993) found that ratings were more reliable and rater drift was eliminated with synthesized auditory anchors for each severity degree for roughness when compared to ratings without the anchor samples. Combined auditory (synthetic) and textual anchors improved rating reliability (Chan & Yiu, 2002). The reliability of listeners trained to use reference matching with anchor samples or paired-comparison tasks were equivalent (Chan & Yiu, 2006). Awan and Lawson (2009) established that combined auditory anchors (natural voice) and text anchors (written descriptions) were most effective in improving interrater reliability, more so than the improvement seen by using auditory anchors alone.

**Listener training.** Studies since 1993 have identified that training improved intra- and interrater reliability (Awan & Lawson, 2009; Chan et al., 2012; Eadie & Baylor, 2006; Ghio et al., 2014; Iwarsson & Reinholt Petersen, 2012). The highest interrater reliability for the GRBAS occurred in a study with experienced expert raters who participated in six days of intensive training in perceptual voice evaluation (Webb et al., 2004). In another successful training program, twenty hours of group training was embedded in a course on voice analysis (Iwarsson & Reinholt Petersen, 2012). That training, which
included imitation, improved the identification and the description of voice qualities and increased interrater agreement. Chan, Li, Law, and Yiu (2012) contrasted providing immediate feedback (expert rating provided after each sample) versus no feedback (summary feedback as an accuracy percentage at the end of each rating session). Both groups improved immediately after training, but only the summary feedback group maintained improvement 5 days after training. Chan et al. (2012) is one of the few studies contrasting ratings immediately after training with subsequent rating sessions. Training methods likely improved reliability by creating common standards between individuals varying in experience and biases, increasing sensitivity to different voice qualities, developing internal standards for voice qualities, and creating consistent definitions within each study.

One aim of understanding the variability in voice perceptual ratings and the influence of training is to develop effective training programs to prepare SLPs, and more specifically SLP Master’s students, to provide accurate vocal perceptual evaluations of patients. This study investigated imitation as another method to improve perceptual training and reduce variability. The next section discusses how imitation and motor learning improved the discrimination and judgment skills of naïve listeners.

The Influence of Motor Learning on Voice Perception

The use of motor production experience to train perceptual skills is not a new concept in speech-language pathology, but limited research has investigated its effectiveness. The approach has face validity and anecdotal support from clinicians and educators, and it is endorsed in a collegiate level textbook (Duffy, 2013). Two studies of motor learning and perception include a training protocol discussed above that included imitation of pathological voices and a study of the influence of singing study on perceptual skills. These studies offer insight on how imitation has been applied for training purposes and how extensive motor training through singing changed participants’ perceptual systems. Both of these ideas are relevant to the development of imitation as a training method for naïve listeners.

Imitating pathological voices. The consensus training protocol developed by Iwarsson and Reinholt Petersen (2012) included imitation as a component of the training, but did not separate the influence of the imitation from the other training components included in the voice course. The training
goal was creating common standards for identifying voice qualities and to solidify students’ internal standards. During the consensus training, participants were provided with voice quality definitions, information on the vocal fold physiology, examples of different severity levels of the qualities studied and imitative vocal exercises. The vocal exercises included practicing sustained vowels ranging from hyper- to hypo-functional, and imitating vocal fry on a vowel and in continuous speech. The researchers indicated that imitation exercises were “for the students to make use of proprioception and thus the concept of creative hearing” (Iwarsson & Reinholt Petersen, 2012, p. 305). The training included the use of reference voice samples, but did not use auditory anchors during the ratings. There was high interrater reliability before the training, and the training increased the level of exact agreement.

**Singing instruction.** DeBoer and Shealy (1995) evaluated the influence of private voice instruction on the perceptual skills of Master’s SLP students and found that 7 weeks of study facilitated positive change in vocal perceptual skills. These perceptual changes were maintained 7 weeks after lessons were discontinued. Twenty graduate students in speech-language pathology were divided into two groups of ten who alternated receiving seven weeks of private voice (singing) lessons. The data suggested that 7 weeks of private voice study was sufficient to develop and establish new vocal-motor patterns and perceptual skills. Evidence for new motor patterns included improvements in the S/Z ratio test, reduced glottal stop production in speech, and improved song performance. Evidence for improved perceptual skills included improved pitch perception and the identification of glottal stops in speech samples. A suggestive correlation was observed between motor patterns and perception: participants who reduced glottal stop production also increased identification of glottal stops in pathological voices. Participants’ subjective evaluations of pathological voice samples also improved. The retention of voice perception and performance skills for 7 weeks after the initial lessons suggested that the new motor production patterns and perceptual changes were stable.

The inclusion of imitation in training for vocal perceptual evaluation supports our hypothesis that imitation could improve perceptual ratings. The hypothesis is further supported by the data that motor experience in singing did result in improved perceptual skills, which suggests a connection between vocal
motor production and perception. These observed changes in perception following vocal motor training with imitation and singing are consistent with models of speech and voice perception, which will be explored in the following section.

**Theoretical Models and Neurological Research**

The hypothesis that motor production experience may affect perception is supported by theories and models from the disciplines of speech and psychology, as well as neurological research on connections between motor production and perception. The results from the two studies about imitation during perceptual training and singing study are consistent with these theories. The following section begins with speech perception models, then a model specific to voice perception, and then concludes with neurological evidence supporting those models.

**Speech perception models.** Speech perception models are included because of the long history of investigating motor speech production and its relationship to perception, beginning with the motor theory of speech production proposed by Liberman and Mattingly in 1957 (Galantucci et al., 2006). The speech perception models are likely applicable to voice because both systems use similar neurological pathways. Unlike noise which is processed bilaterally by both hemispheres, both speech and voice perceptual systems are lateralized to the left hemisphere for perception and production (Chang et al., 2009; Hickok & Poeppel, 2000). Moreover, voicing is an essential component of speech perception for distinguishing between numerous voiced/unvoiced phoneme pairs, such as /b/ and /p/. Voice onset time (VOT), which is the time between when a phoneme begins and when voicing begins, defines whether a voiced (/b/) or unvoiced (/p/) phoneme is perceived. The speech perceptual system identifies millisecond variations in VOT during speech production (Goldrick & Blumstein, 2006). Articulation research also suggested that the variables interacting during the production of speech may actually change the underlying structure of the perceptual system (Baese-Berk & Goldrick, 2009). Linebaugh and Roche (2013) found that articulation training with unfamiliar phonemes for second-language learners of English improved the auditory discrimination between those phonemes. Diehl, Lotto, and Holt explained observed connections between speech production and perception with two general ideas: “production
follows perception, and perception follows production” (2004, p. 167). This potential to change the perceptual system is the premise this study is testing with the use of imitation for training vocal perceptual skills. Furthermore, the one voice perception model to be discussed (Hutchins & Moreno, 2013) draws heavily on speech models in form and structure (Hickok & Poeppel, 2000). The hypothesized overlap between the speech and voice perceptual systems supports reviewing speech models or possible mechanisms for imitation to influence the perception of voice quality in our naïve listeners.

The motor theory of speech perception developed by Liberman and Mattingly (1985) connects the perception of speech to its motor production. The first claim is that “to perceive an utterance, is to perceive a specific pattern of intended gestures” (1985, p. 3). The second claim is that speech processing is special compared to other sensory processing, and the third claim is that speech perception involves access to the speech motor system. The model proposes that speech perception is an interpretation of vocal tract motor movements and that the motor cortex is accessed during this process. This connection between motor production and perception relates to this study. Galantucci et al. (2006) presented evidence for motor involvement in speech perception, perception-motor connections in other communication systems, motor connections to perception of motion, and motor system recruitment for perception. Based on these results, those researchers concluded that Liberman’s contention that speech perception accesses the motor system was likely valid and went further to state that evidence supported a general connection between motor experience and perception. This general connection could apply to voice perception.

Hickok and Poeppel (2000) proposed the dual-stream model of speech perception (Figure 2), which divides speech perception into two parallel neurological streams. The first stream processes comprehension (ventral stream/auditory-conceptual interface), and the second stream processes production (dorsal stream/auditory-motor interface). Auditory input is routed via the ventral stream to Wernicke’s area at the temporal-parietal-occipital junction for conceptual interpretation. The dorsal stream routes the signal to the inferior parietal and frontal lobes for sensory and motor integration. The theory splits speech processing into two independent systems, rather than processing motor and cognitive
information serially as in the motor theory. To explain evidence that perception can affect production, this theory proposes that the two streams will connect speech comprehension and production information after the initial processing is complete (dotted line in Figure 2). The connection between the ventral and dorsal streams is the location where imitation experience could influence the perception of voice quality.

![Figure 2 Dual-stream model of speech perception. “A simple model of the cortical networks supporting speech perception and related language functions. The dashed line indicates the possibility of additional, non-parietal auditory-motor interface networks.” Reprinted from Trends in Cognitive Sciences, Vol 4/4, G. Hickok and D. Poeppel, Towards a functional neuroanatomy of speech perception, Page 132, Copyright (2000), with permission from Elseviers.]

**Voice perception model.** Building on the previously discussed models of speech perception and production, Hutchins and Moreno (2013) proposed a dual-route model of voice perception called the linked dual representation model (Figure 3). Low-level perception of the auditory signal is divided and directed to a vocal-motor representation and a symbolic (cognitive) representation. The symbolic representation feeds forward to the vocal-motor representation and thus cognitive knowledge has a secondary influence on production. The feedback loop from the vocal motor representation to low-level perception would include both auditory feedback and efferent feedback from the motor production and could be a route for motor experience to influence perception. This dual linked model accounts for some of the broad correlations found between vocal perception and production abilities. These correlations
include results that individuals who have poor pitch perception tend to be worse singers and that most people’s imitations will generally line up with their perceptions, though not perfectly (Hutchins & Moreno, 2013). The model is intended to widely apply to vocal production and perception beyond the area of pitch perception, but has not been researched for voice quality discrimination. This study considers the validity of this vocal perceptual model beyond pitch to the area of voice quality.

Figure 3 The linked dual representation model. The curved line represents a feedback loop from the vocal-motor representation to low-level perception (Hutchins & Moreno, 2013). This figure is reprinted under the terms of the Creative Commons Attribution License (CC BY).

Feedback mechanisms in the perception system. Both the dual-stream model of speech perception and the dual linked representation model of voice perception include feedback mechanisms linking perception and production. It is bidirectional for the model of speech perception and unidirectional from perception to production for the model of voice perception. Many biological systems include feedback, and there are numerous forms of feedback in the vocalization process (see also the following section). Haskell (1987) identified two major forms of vocal feedback, aural and tactile-proprioceptive, that correspond to the feedback loop from the vocal-motor production to the low-level auditory perception in the dual linked representation model. Conscious awareness of motor production sensations can become an active part of the self-perceptions of one’s voice (Haskell, 1987). Singers
monitor these physical sensations and impressions to train and evaluate their productions (Hines, 1984). The premise of this study relies on the presence of a mechanism connecting perception and production, and research into mirror neurons provides evidence of a neurological connection between these processes.

**Neurological support for theoretical models of perception.** Mirror neurons are motor neurons in the brain that fire when a motor action is completed and also when that same motor action is observed or heard. General evidence related to mirror neurons supports the presence of a bidirectional connection between perception and production. These connections may facilitate the ability of motor experience to influence conscious and unconscious perceptual abilities. Research supporting the connection between production and perception in the areas of manual manipulation, auditory perception, and dance perception follow. These connections between motor production and perception may be relevant for vocal imitation and perception of voice quality.

The motor neurons that produce voice and speech are active when perceiving speech and vocalizations. Kohler et al. (2002) established in monkeys that auditory sounds related to an action will activate the motor neurons necessary to produce that action. Researchers have also identified motor neuron activity in humans during speech listening and listening to other human sounds (Aglioti & Pazzaglia, 2010; Chang et al., 2009). Lévêque, Muggleton, Steward & Schön (2013) further tested this finding with transcranial magnetic stimulation (TMS) and observed involvement of the premotor laryngeal areas in voice perception.

In addition to the mirror neurons in the motor cortex, there may also be a connection between voice production experience and auditory processing centers. Zarate and Zatorre (2008) observed that the dorsal premotor cortex was involved in processing and managing manipulated auditory feedback provided to singers. This may be a sensory-motor interface for perception and production. Kleber et al. (2010) found that during a listening task, experienced singers recruited a larger area of the “singing network,” which includes the auditory cortex, and also activated the somatosensory cortex associated with the larynx and mouth. The experienced singers engaged different cortical systems, and the response was more automatic than in less experienced singers.
Similarly, research comparing elite dancers of two different styles observed increased neural activation in individuals with specific motor production experience. Functional magnetic resonance (fMRI) studies established that in elite dancers of two different styles, ballet and capoeira, motor neurons were more active when watching the style the subject performed than the other style of dance (Calvo-Merino, Glaser, Grèzes, Passingham, & Haggard, 2005). The subjects responded more strongly to the movements in their physical repertoire. Additionally, the responses of both elite groups were unique from the responses of non-dancers. The researchers concluded that the brain’s response was dependent on the individual’s repertoire of skilled, learned motor behavior.

Neurological studies identified connections between the motor production and perception systems. For the purposes of evaluating dysphonic voices, this information suggests that a listener with more motor experience producing dysphonic voice qualities may have more neurological activation when listening to pathological voices.

**Summary.** SLPs are likely to encounter voice clients, yet training and preparation varies widely. Training is important, but developing subjective, perceptual skills takes time and is difficult: even experts vary in their accuracy and agreement when evaluating voice quality. We know that extensive training, anchor samples, and delayed feedback can improve rater accuracy. Theoretical models from speech and voice perception support the idea that imitation could provide an additional means of improving reliability. The motor theory of speech perception predicts a serial connection between the motor and perceptual systems, and the dual-stream model of speech perception and the linked dual representational model of voice perception propose two routes for processing comprehension and production that later reconnect. Neurological research in humans suggests there is a connection between voice production and perception. The information presented included a rationale for proposing that imitating pathological voices could influence the perception of voice quality. The purpose of this study was to assess changes in accuracy and confidence levels associated with the addition of imitation to a perceptual voice assessment training protocol.
Methodology

This study was structured as a comparison between the pretest and post-test perceptual voice evaluations of pathological voices completed by two groups of naïve listeners. One group received didactic training in vocal perception alone, and the other group received didactic training with motor experience (i.e., imitation of pathological voice qualities).

Study Participants

Naïve listeners. The naïve listeners were seven undergraduate university students, between the ages of 18 and 37, with no coursework in voice or neurological disorders of speech and no formal study or collegiate experience with singing and acting. Participants were native speakers of American English with reported normal hearing. They were recruited from undergraduate students in the University of Iowa Communication Sciences and Disorders Department via posters in the department, a departmental email, class presentations and a Facebook post to the Iowa chapter of the National Student Speech-Hearing-Language Association. An all-campus mass email message was sent to UI undergraduates. The recruitment and research procedures were approved by The University of Iowa’s Institutional Review Board. The undergraduate participants were divided into two groups: a no-imitation control group (N = 3) that received only didactic training, and an imitation group (N = 4) that received didactic training and vocal motor experience imitating pathological voice qualities. Participants were assigned randomly to the groups at a 1:1 ratio based on order of enrollment.

Expert listeners. There are no correct answers in a perceptual rating task; therefore, it was necessary to recruit expert listeners (N = 3) to develop a gold standard rating for determining the accuracy of the perceptual ratings by the naïve listeners. Three speech-language pathologists with at least eighteen years’ experience evaluating voice disorders and familiarity with the GRBAS protocol were recruited. Their data were used to develop a summary expert rating and select anchor samples. The expert listeners were native speakers of American English. Two reported normal hearing and one reported tinnitus with an unconfirmed possible high frequency hearing loss. Expert listeners were recruited from voice professionals in Iowa and Minnesota through an email message.
Demographic data were collected for all participants, including a vocal self-rating, experience with singing, experience in music, and related information.

**Voice samples.** Twenty-five voice samples were rated by the experts. Two samples were dropped and twenty-three of the voice samples were used for the naïve listener training: six samples for training sessions (training set), six samples for the homework (homework set), six samples for the pre- and post-tests (test set), and five samples for anchor samples (anchor set) that were available and used in all sessions except the homework. The training and test sets included samples of patients with voice disorders and normal speakers. The anchor sample set was exclusively pathological voices. The voice disorders included membranous vocal fold lesions, muscle tension dysphonia, vocal fold paralysis, spasmodic dysphonia, and/or tremor. The initial set of voice samples were 52% male and 48% female, and this distribution was maintained after two voice samples were dropped. Each voice sample included a sustained vowel (/a/) for approximately 4 seconds and one or more repeated sentences (e.g., “We were away a year ago,” “We eat eels every Easter,” etc.). Voice samples were drawn from samples collected at the University of Iowa Hospitals and Clinics (UIHC) as part of laryngeal function studies during voice evaluations between January 2009 and August 2015. A waiver of consent was granted for these participants.

Speech samples were recorded in the Voice Clinic’s procedure rooms in the Department of Otolaryngology at UIHC. Participants were recorded via a headset microphone approximately 4 inches from the participant’s mouth and off center. Files were recorded through the KayPENTAX Computerized Speech Lab 4500 (CSL) and originally saved as CSL NSP files. Recording volume was monitored and samples rerecorded if necessary to avoid clipping. For this study, speech sample files were converted to .WAV files via CSL.

**Expert Listener Sessions, Anchor Samples and Summary Ratings**

**Expert listener sessions.** Expert listeners participated in an initial session of 15 to 20 minutes with the principal investigator (PI) where participants were screened, consent was obtained and demographic information was collected. The voice samples were rated online in a second session at a
time and location of the subject’s choice. The self-paced, online PowerPoint 2013 presentation was available through Microsoft OneDrive for Business. Participants were directed to listen to both the sustained vowel and sentence recordings for all samples and to rate the samples as they proceeded through the presentation. Voice quality ratings were handwritten on a hard copy data collection sheet, and a separate reference sheet with voice quality definitions was provided with the data sheet. The expert participants could listen to samples as many times as they wished, stop the session and return later, and compare samples as they were making the ratings and anchor sample recommendations. Instructions were not provided regarding the use of headphones or speakers for listening.

The PowerPoint presentation began with a review of the 1981 GRBAS protocol and definitions of the five voice qualities from Hirano (1981). The GIRBAS protocol (Dejonckere et al., 1996) was introduced and instability defined. Expert listeners rated 25 voice samples on the GIRBAS perceptual rating scale, with five samples repeated to assess intrarater reliability, for a total of 30 voice sample ratings. Two of the three expert SLPs provided confidence ratings to indicate their confidence in their GIRBAS ratings. The confidence scale was from one to five with one indicating a high level of confidence and five indicating a low level of confidence. Expert listeners were invited to make comments about any of the samples during the ratings, make notes about possible anchor samples during the initial voice ratings, and answer two questions at the end of the rating session about the composition of voice samples and how they rate voices.

**Selecting anchor samples.** Expert listeners recommended appropriate anchor samples for the naïve listener training. They were informed that having an external reference (“anchor stimulus”) improved reliability of perceptual ratings (Gerratt et al., 1993) and were directed to select “clear” samples of the different voice qualities. After completing all 30 ratings, expert listeners were prompted to identify clear samples to serve as anchors for instability, roughness, breathiness, asthenia and strain and to provide reasons for the selection of those samples. Anchor samples were selected from the voice samples suggested by the experts. If two to three expert listeners identified a unique sample as a strong anchor for voice quality, it became the anchor sample. For roughness, two different samples were selected by two
experts, and the PI selected between those samples. One sample was selected by the experts for both breathiness and asthenia anchors. That sample was discarded as an anchor because it could potentially confuse naïve listeners. The principal investigator selected the anchor samples for breathiness and asthenia from the other samples recommended by one expert listener as exemplars of these qualities.

**Establishing summary expert ratings.** The summary expert ratings served as a gold standard for evaluating naïve listener ratings. The summary ratings were developed from the individual expert ratings for each voice quality by an agreement method. If two or three experts agreed on the severity rating for a specific voice quality (Grade, Instability, Roughness, Breathiness, Asthenia, and Strain), then that severity rating became the expert rating. When none of the experts agreed on the severity rating, the ratings were split into two groups: high (2-3) and low (0-1). In the group (high or low) with two scores, the medial score (either 2 or 1) was selected as the expert rating (i.e., if two experts selected high scores of two and three, the expert rating was a two. If two experts selected low scores of zero and one, then the expert rating was a one). In the five samples rated twice, all six ratings were considered to determine the expert rating. If the expert ratings from the first and second ratings were the same, that score became the expert rating. If the first and second ratings were different, scores with four matching ratings became the expert rating. If the first and second ratings did not agree and there were not four matching ratings, then the second rating became the expert rating.

A consistency rating for each summary expert rating was calculated to identify how consistent the ratings were across the raters. Points from one to three were assigned for each of the six voice qualities based on the number of expert severity ratings that agreed, for a total possible of 18 points per sample. A high score indicated a high level of consistency in the expert ratings, and a low score indicated less consistency in expert ratings. Two voice samples with consistency ratings of 10 points or lower were excluded from the naïve listener sessions. A similar scale of 36 points was used for the five samples rated twice, and all five voice samples received confidence ratings of 21 points or higher and were included. Voice samples in each naïve listener session included a mixture of expert consistency ratings and
differing severity levels for grade, plus a variety of voice qualities and vocal pathologies. The anchor samples all met the inclusion criteria of a consistency score greater than 10.

**Naïve Listener Sessions**

Naïve listeners participated in four sessions of 15 to 60 minutes: three in-person with the principal investigator and one online as homework. Participants completed the in-person sessions in an office at the Wendell Johnson Speech and Hearing Center. Sessions with the PI were separated by at least a week and were completed over the course of 3 to 4 weeks. Participants completed all in-person sessions on a PC using headphones with a microphone. The homework session was completed online at a location of the subject’s choosing with no instructions provided regarding headphones or speakers.

**Session protocols.** In each naïve listener session, subjects rated 6 voice samples with the GIRBAS, gave a confidence rating, and made any additional written observations. Participants were discouraged from directly comparing samples, but could use the anchor samples as references. Audio anchor samples for instability, roughness, breathiness, asthenia and strain were provided each time a voice sample was rated during the pretest, training, and post-test. Participants could listen to the anchor samples as often as they wished, but were not required to listen to the anchors. Audio anchors were not available for the homework. A reference sheet with the voice quality definitions presented in the training was also provided at each session and made available for the homework. Subjects were prompted to listen to all vocal samples three times with or without imitation and to move through the training at their own pace. All users were instructed to listen to both the sustained vowels and the sentences, but to focus on the sentences for the ratings. The imitation group was directed to “go for the feeling,” rather than be concerned about exactly imitating the samples. Samples could be repeated and imitated as much as a subject wished. The length of the in-person sessions was documented. Additional information collected included questions asked by the subject and PI responses, side comments, whether the subject appeared to listen or reported listening three times, whether imitation was completed, and other notes. Accuracy of the imitation was not noted or addressed. Audio recordings were made of the imitation group for the training and post-test.
Session activities. In session one, participants completed screening, consent and intake documents and completed the pretest. The pretest was a PowerPoint 2013 presentation with embedded audio files that briefly introduced voice quality definitions and the GIRBAS protocol. Six voice samples were rated, and the pretest took approximately 15 minutes to complete. The second session was a PowerPoint 2013 training module modeled after online training materials included with Motor Speech Disorders: Substrates, Differential Diagnoses, and Management by Joseph R. Duffy (2013) and took approximately 45 minutes to complete. The training module included more detailed information on the GIRBAS protocol, guidelines for developing perceptual listening skills, and descriptions of each voice quality with audio anchor samples that typified each of those qualities. Participants rated six new voices. After completing the ratings, they listened again to the voice samples they rated and checked their answers against the summary expert ratings. As this was a perceptual evaluation, without definitive right or wrong answers, participants were reminded that the expert ratings were provided as references, not as the absolute correct answers (Yamaguchi et al., 2003). The session ended with four to five written open-ended questions for the participants about the training and their learning.

A PowerPoint 2013 homework presentation and a downloadable answer sheet were available between the training session and the final in-person session via a web link to OneDrive for Business. Subjects rated six additional voice samples and followed the same procedures as during the training session. Anchor samples were not available for the homework because of a technical glitch. Participants could submit their completed homework before the final session, and then receive a link to a presentation with the summary expert ratings. The final session was a post-test with the same voice samples in the same order as the pretest. This session concluded with four questions about learning that occurred through the project and what changes the subjects would suggest.

Participants in the no-imitation group received didactic training only and served as a control group. The imitation group received didactic training and imitated the voice samples. During the training, homework and post-test, the imitation group was directed to imitate each voice sample three times and mark on the answer sheet that they completed the imitations. Compliance with these directions
was tracked by the PI for the in-person sessions, and a compliance rating for listening or imitation was requested from the participant after the post-test.

Data Analysis

The study examined the role of training naïve listeners’ ability to discriminate between pathological voice qualities during a vocal perceptual evaluation, and the influence of motor experience (imitation) on those same vocal perceptual skills. The manipulated variable was the type of training with one group of naïve listeners receiving didactic training and one group of naïve listeners receiving didactic training and additional motor (imitation) experience. Measured variables included intra- and interrater reliability of expert listeners, interrater reliability between the naïve listener groups compared to the summary expert ratings, and raters’ confidence levels.

Demographic data, such as age and experience level, were analyzed using basic descriptive statistics: mean, range, and standard deviation. Additional binomial demographic data, like male and female, were analyzed with percentages.
Severity ratings for grade, instability, roughness, breathiness, asthenia, and strain were collected for each voice sample. Intrarater reliability for the expert voice quality ratings was calculated using Cohen’s kappa. This statistic compares the number of times the first rating matched the second rating to times when the ratings did not match from the first rating to the second rating. Agreement levels of greater than 80% are considered very good and 60-80% good. The interrater agreement for the experts was calculated by counting the number of scores with the exact match (three experts), agreement (two experts), and no agreement and calculating the percentages. General agreement was calculated by adding the number of scores with two to three experts agreeing (exact matches and agreement).

The hypothesis addressed the accuracy of naïve listener voice quality ratings for the entire group and for the imitation versus control group. Accuracy was calculated by calculating the absolute value of the difference between each naïve listener’s ratings and the expert ratings. This became the divergence score. These divergence scores were averaged across individual voice qualities by subject or by session and were averaged across all voice qualities and samples by session. The maximum possible divergence is 3 (i.e., the difference between a summary expert rating of 0 and a naïve listener rating of 3). The minimum divergence is 0 (i.e., the naïve listener matched the expert score). Cohen’s d (i.e., pooled standard deviation $\frac{M_{post} - M_{pre}}{0.5(SD_{post} - SD_{pre})}$) was calculated to compare a standardized difference between scores and determine the effect size of the difference. An effect size of 0.8 or more is large, 0.5 medium, and 0.2 small.

Accuracy data were graphed with voice qualities on the x-axis and the divergence from the expert standard on the y-axis. Divergence scores closer to the x-axis are more accurate, and scores that are further from the x-axis are less accurate. Comparisons were made between pretest and post-test scores and between the post-test scores of the imitation and control groups. Changes in divergence across all four sessions were graphed with the session on the x-axis and divergence from the expert standard on the y-axis, and voice quality was indicated by marker shape and color. The average divergence by subject
and session was presented on a single graph with subject and session on the x-axis. Another group of graphs presented this information by individual participant with voice qualities on the x-axis, divergence on the y-axis, and different colored lines indicating the session.

Confidence levels (1 = very confident, 5 = not at all confident) were collected for each voice rating and descriptive statistics (mean, range, range size) were calculated for the experts and the naïve listeners. Because the pretest confidence levels were different between the imitation and control groups, normalized gain scores were calculated to compare post-test confidence levels of the imitation and control groups (Ebbels, Van Der Lely, & Dockrell, 2007). This score compares the actual gain with the potential for gain for each of the groups [i.e., (Post-test score – Pretest score)/(1-Pretest score)] and is independent from the confidence level scale. The five-point confidence score was converted into a percentage, and then the normalized gain was calculated.

A power analysis was conducted to establish the sample size necessary for statistically significant results. The assumptions were alpha 0.05, power 0.8, effect size of 0.1, and standard deviations for both groups of 0.08. The estimated required sample size was 22 (11 in the imitation group and 11 in the control group).
Results

Demographic information for all subjects, and then results for expert and naïve listeners, are presented.

Participant Group Demographics

Tables for the expert listeners are presented first, followed by all naïve listeners, and then imitation and control groups.

The expert listeners had an average age of 58.7 years (standard deviation ± 4.7 years) and an average of 25 years’ experience (standard deviation ± 6.6 years) experience rating voices. The average voice quality self-rating was 2 (standard deviation ± 1) on a scale of 1 to 6, with one being excellent. All experts had at least five years of instrumental music experience. See Table 1 below.

<table>
<thead>
<tr>
<th>Table 1 Expert Listener Demographics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mean</strong></td>
</tr>
<tr>
<td>Age</td>
</tr>
<tr>
<td>Experience</td>
</tr>
<tr>
<td>Voice Quality Self-Rating</td>
</tr>
<tr>
<td>Normal Hearing</td>
</tr>
<tr>
<td>Male/Female</td>
</tr>
<tr>
<td>Instrumental/None</td>
</tr>
</tbody>
</table>

<sup>a</sup> tinnitus and suspected high frequency hearing loss.  
<sup>b</sup> > 4 years.

The average age of the naïve listeners was 24.9 (standard deviation ± 7.08) with an average voice quality self-rating of 2.4 (standard deviation ± 1.26). The group was 100% female, 86% communication sciences and disorders majors (CSD) and 43% CSD seniors. Eighty-six percent (86%) of the naïve listeners had instrumental music training of between one and eight years. All naïve listeners had reported normal hearing and met all inclusion criteria (Table 2).
Table 2 Naïve Listener Demographics

<table>
<thead>
<tr>
<th></th>
<th>All naïve listeners</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Range</td>
<td>SD</td>
</tr>
<tr>
<td>Age:</td>
<td>24.9</td>
<td>19-37</td>
<td>7.08</td>
</tr>
<tr>
<td>Voice Self-Rating:</td>
<td>2.4</td>
<td>1-4</td>
<td>1.26</td>
</tr>
<tr>
<td>Male/Female</td>
<td>0%</td>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>Instrumental/None</td>
<td>86% a</td>
<td></td>
<td>14%</td>
</tr>
<tr>
<td>CSD Major/Other</td>
<td>86%</td>
<td></td>
<td>14%</td>
</tr>
<tr>
<td>CSD Senior/Other</td>
<td>43%</td>
<td></td>
<td>57%</td>
</tr>
</tbody>
</table>

a 1-8 years.

There were differences between the imitation group and the control group. The average age for the imitation group was 27.5 years (standard deviation ± 8.9 years) and for the control group was 21.3 years (standard deviation ± 0.58). The imitation group voice quality self-rating averaged 2.75 (standard deviation ± 1.4) and the control group averaged 2 (standard deviation ± 0.71). Seventy-five percent (75%) of the imitation group had instrumental music experience ranging from 1 to 5 years. One hundred percent (100%) of the control group had instrumental music experience ranging from 3 to 8 years. The imitation group was 75% CSD majors and 25% CSD seniors, while the control group was 100% CSD majors and 66% CSD seniors (Table 3).
Table 3 Naive Listener Demographics by Experimental Group

<table>
<thead>
<tr>
<th></th>
<th>Age (years)</th>
<th>Voice Self-Rating (1-6)</th>
<th>Sex</th>
<th>Instrument (years)</th>
<th>CSD Major</th>
<th>CSD Senior</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Imitation Group</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subject 1</td>
<td>19</td>
<td>3</td>
<td>F</td>
<td>0</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Subject 3</td>
<td>33</td>
<td>1</td>
<td>F</td>
<td>5</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Subject 5</td>
<td>37</td>
<td>3</td>
<td>F</td>
<td>3</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Subject 7</td>
<td>21</td>
<td>4</td>
<td>F</td>
<td>1</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td><strong>Mean/%</strong></td>
<td>27.5</td>
<td>2.75</td>
<td>100%</td>
<td>2.25</td>
<td>75%</td>
<td>25%</td>
</tr>
<tr>
<td><strong>SD</strong></td>
<td>8.9</td>
<td>1.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Control Group</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subject 2</td>
<td>21</td>
<td>2</td>
<td>F</td>
<td>5</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Subject 4</td>
<td>22</td>
<td>3</td>
<td>F</td>
<td>3</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Subject 6</td>
<td>21</td>
<td>1</td>
<td>F</td>
<td>8</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td><strong>Mean/%</strong></td>
<td>21.3</td>
<td>2</td>
<td>100%</td>
<td>5.3</td>
<td>100%</td>
<td>66%</td>
</tr>
<tr>
<td><strong>SD</strong></td>
<td>0.58</td>
<td>0.71</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Expert Results**

**Interrater agreement.** Expert agreement levels were calculated by determining the percentage of time the experts had the same severity rating on individual voice qualities for each sample. The experts initially rated 25 unique samples with six voice quality severity ratings for a total of 150 ratings. Three experts agreed on 64 ratings, two experts on 69, and no experts on 17 for percentages of 43%, 46% and 11%, respectively, and a general agreement of 89%. Agreement levels are reported for all samples, samples for the naïve listener trainings, pretest and post-test samples, practice samples rated during the training session and homework, and the anchor samples. The rows in Table 4 are the number of experts that agreed and the columns are the different voice sample groups. The last row includes the percentage of general agreement between two or three experts.

In the initial expert GIRBAS ratings, two to three experts agreed on the severity rating 89% of the time for individual voice qualities. The agreement level increased to 92% after two voice samples were dropped for poor agreement. Agreement levels for samples selected for different sections of the study varied from 89% (training and homework practice samples) to 90% (anchors) and to 94% (pre-/post-test).
Exact agreement (three experts agreed) varied from 33% to 53%, with the pre-/post-test voice sample set at the low-end (33%) and the anchor set at the high-end (53%). See Table 4.

**Table 4 Expert Listener Agreement for Voice Samples**

<table>
<thead>
<tr>
<th># of Experts Agreeing</th>
<th>All Samples (N=25)</th>
<th>Training Samples (N=23)</th>
<th>Pre/Post-test (N=6)</th>
<th>Training Practice (N=6)</th>
<th>Homework Practice (N=6)</th>
<th>Anchors (N=5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>43%</td>
<td>46%</td>
<td>33%</td>
<td>44%</td>
<td>44%</td>
<td>53%</td>
</tr>
<tr>
<td>2</td>
<td>46%</td>
<td>47%</td>
<td>61%</td>
<td>44%</td>
<td>44%</td>
<td>37%</td>
</tr>
<tr>
<td>0</td>
<td>11%</td>
<td>7%</td>
<td>6%</td>
<td>11%</td>
<td>11%</td>
<td>10%</td>
</tr>
</tbody>
</table>

**General Agreement (2-3 experts)** | 89% | 93% | 94% | 89% | 89% | 90%

**Intrarater agreement.** The ratings for the five voice samples that were rated twice by each expert were compared using Cohen’s kappa. Cohen’s kappa showed good agreement (M = 80%) with a range of 73% to 90% for the individual experts.

**Confidence levels.** Two out of three experts provided confidence levels for the voice samples (1 = very confident, 5 = not at all confident). The confidence levels were averaged across all the samples and for the samples used in the pretest and post-test. The average expert confidence level was 2.0 with a range of 1-3. The confidence level for the pre- and post-test set was 2.50 with a range of 2–3.

**Anchor samples.** Eighteen different samples were suggested as possible anchor samples. This included six samples that 2–3 experts agreed on (Table 5). Final anchor samples were selected according to the criteria specified in the methodology.

**Table 5 Expert Listener Agreement for Anchor Samples**

<table>
<thead>
<tr>
<th># of Experts Agreeing</th>
<th>Instability</th>
<th>Roughness</th>
<th>Breathiness</th>
<th>Asthenia</th>
<th>Strain</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td>1&lt;sup&gt;b&lt;/sup&gt;</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2&lt;sup&gt;a&lt;/sup&gt;</td>
<td></td>
<td>1&lt;sup&gt;b&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

<sup>a</sup> PI selected between two suggested anchors. <sup>b</sup> The same sample was selected for both of these qualities. Breathiness and asthenia anchor samples were selected by the PI from other suggested samples.
**Narrative questions.** The expert listeners were asked two questions for each sample during the GIRBAS ratings (i.e., Additional observations? Any reason to exclude this sample?), why they selected specific anchor samples, and two narrative questions at the end of the rating session (i.e., Is this a good mix of voice samples for training undergraduate listeners? What is missing? and Any additional comments or suggestions?).

Comments during the ratings included basic observations about the samples (e.g., “diplophonia,” “normal female voice,” “very harsh quality”) and no samples were suggested for exclusion. Comments about anchor sample selection included identifying the diagnosis (for instability “tremor, S[pasmotic] D[yshonia]”), and indicating the quality was isolated (“primary characteristic is roughness,” “less influence of other qualities”).

Responses to the two narrative questions at the end of the rating session were as follows. The experts agreed that the mix of voice samples was good, but one expert recommended adding some less severe samples. For additional comments, experts mentioned feeling less confident with instability since it was not part of the typical GRBAS scale, that voice qualities often occur together and that anchors may need to reflect this, and that one rater focused on the sentence for the ratings. Several procedural and formatting suggestions were also made.

**Naïve Listener Results**

Naïve listener results are presented to address the question of whether training improved accuracy and confidence levels, and whether imitation improved accuracy and confidence levels when compared to the control group.

**Change in accuracy of voice quality rating with training.** The results indicated that training did improve accuracy of ratings. The average divergence by voice quality is presented in Figure 5 with the different voice qualities on the x-axis and the total divergence on the y-axis.

The average divergence was 0.63 (standard deviation ± 0.18) in the pre-test was and 0.50 (standard deviation ± 0.13) in the post-test with a good effect size (Cohen’s d = 0.83). Average
divergence for the individual voice qualities improved for all qualities, except breathiness, which was approximately the same for both ratings. The results for individual voice qualities are shown in Figure 5.

![Divergence Pretest to Post-test](image)

**Figure 5** Naïve listener divergence pretest to post-test by voice quality. Accuracy changes were calculated by taking the difference between the naïve listener severity ratings and the summary expert ratings for each voice sample and then averaging the absolute values of these differences for the individual voice qualities.

**Group trends across sessions.** Changes in average divergence in voice quality ratings were noticed across all naïve listener sessions (Figure 6). The x-axis is the session (pretest through post-test) and the y-axis remains average divergence from the expert rating. The different voice qualities are indicated by marker shape and color. Trend lines for each voice quality are also included and are indicated by line style and color.

The average divergence for the pretest and post-test were reported previously and are not repeated here. The average divergence on the training was 0.48 (standard deviation ± 0.24) with a medium effect size (Cohen’s d = 0.76). When the training was compared to the pretest, the average divergence improved for grade, breathiness, asthenia and strain, and remained essentially constant for instability and roughness. The average divergence on the homework was 0.60 (standard deviation ± 0.24) with no effect size (Cohen’s d = 0.11). When the homework was compared to the pretest, the average divergence for the
individual voice qualities improved for grade, breathiness, and strain, remained constant for instability, and was worse for roughness and asthenia. Asthenia appeared to be an outlier on the homework.

The trend lines indicate the overall changes throughout the training protocol, and descending lines indicate improvement in accuracy level. The largest improvement was for strain, with slight improvements for grade, instability, roughness, and asthenia, and a slight decrease in accuracy for breathiness (Figure 6).

![Divergence in Voice Quality Ratings across Sessions](image)

*Figure 6* Naïve listener divergence by voice quality and session. Plotted points indicate voice quality average divergence for a specific session. Trend lines in matching colors identify trends across the study. Descending lines indicate the ratings increased in accuracy and the ascending line indicates the ratings decreased in accuracy.

**Individual trends across sessions.** Data for individual subjects showed similar overall trends. Figure 7 plots average session divergence scores grouped by subject and session (x-axis). Five subjects improved from the pretest to the post-test, one subject was the same and one declined. Similarly, six subjects improved or stayed the same from pretest to the training session. On the homework, six subjects decreased accuracy slightly, but did not decrease to pretest levels. One subject improved on the homework. See Figure 7.
Figure 7 Naïve listener average divergence by subject and session. The average divergence from the summary expert rating across all voice samples was calculated individually for each subject and each session.

When reviewed by voice quality and individual subject, the divergence in voice quality rating scores changed in a variety of ways for subjects across the four sessions (Figure 8). Sessions are indicated by the shaded area or dashed line: pretest (shaded), training (long-dashed line), homework (dotted line) and post-test (solid line).

For most qualities and for most subjects, there was an improvement from pretest (shaded area) to post-test (solid line) which is indicated by the post-test line being within the shaded area. One general trend was that the homework session was the most challenging, especially asthenia. Another trend was
that six out of seven subjects scored worse than the pretest on at least one quality in one session (indicated by dotted, dashed, or solid lines above the shaded area). Only subject two was equal to or better than the pretest scores for all subsequent sessions (indicated by all lines occurring within the shaded area). Subjects three and four only had 1–2 scores worse than the pretest, while at the other extreme subjects one, five, six, and seven had 6–9 scores (out of a possible 18) worse than the pretest. Some subjects (subjects one, three and six) were more consistent in their ratings (indicated by the ratings for the different sessions appearing close together). The other subjects’ ratings (subjects two, four, five and seven) varied widely from session to session. There are no clear trends for the imitation group versus the control group.
Figure 8 Naïve listener divergence pretest to post-test by voice quality and subject. The shaded area indicates the pretest average divergence scores with the various lines indicating scores for the training (long dash), homework (dotted) and post-test (solid).
Changes in confidence level with training. Confidence levels (1 = very confident, 5 = not at all confident) were collected for each voice rating. The average pretest confidence level was 2.73 (standard deviation ± 0.58) and post-test was 2.36 (standard deviation ± 0.23), an improvement of 0.37 with a large effect size (Cohen’s d = 0.91). The size of the range compressed from the full scale (1-5) in the pretest to three scale degrees (1-3) in the post-test (Table 6).

<table>
<thead>
<tr>
<th>Experts (N = 2)</th>
<th>Mean</th>
<th>Range</th>
<th>Range Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Undergraduates Pretest</td>
<td>2.73</td>
<td>1-5</td>
<td>5.00</td>
</tr>
<tr>
<td>Undergraduates Post-test</td>
<td>2.36</td>
<td>1-3</td>
<td>3.00</td>
</tr>
</tbody>
</table>

Change in accuracy with the addition of imitation. The results of this small pilot study indicated that imitation did not improve accuracy of ratings. The average post-test divergence in voice quality rating was 0.50 (standard deviation ± 0.15) for the control group and 0.49 (standard ± 0.12) for the imitation group with no effect size (Cohen’s d = 0.04). There were small effect sizes for three voice qualities, grade (Cohen’s d = 0.28), asthenia (Cohen’s d = 0.25), and strain (Cohen’s d = 0.25), and no effect size for instability, roughness, and breathiness (Figure 9).
Figure 9 Naïve listener post-test divergence by voice quality and experimental group. Divergence was calculated by the method described for Figure 5.

Change in confidence level with the addition of imitation. The average confidence level on a scale from 1 to 5 (1 = very confident, 5 = not at all confident) was calculated for both naïve listener groups for the pretest and post-test. Because the pretest confidence levels were different between the imitation (2.92) and control groups (2.52), normalized gain scores were calculated to compare post-test confidence levels of the imitation and control groups (Ebbels et al., 2007).

The average confidence levels from the pretest to the post-test for the imitation group improved by 0.42 to 2.50 (standard deviation ± 0.27) and 0.36 to 2.17 (standard deviation ± 0.26) for the control group with a small effect size (Cohen’s d = 0.23). The range for both groups reduced to 1–3 on the post-test. Normalized gain was larger for the imitation group (0.20) than for the control group (0.15). See Table 7.
Table 7 Naïve Listener Post-test Confidence Levels

<table>
<thead>
<tr>
<th></th>
<th>Pretest</th>
<th>Post-test</th>
<th>Difference</th>
<th>Normalized Gain Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Mean</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Imitation Group</td>
<td>2.92</td>
<td>2.50</td>
<td>0.42</td>
<td>0.20</td>
</tr>
<tr>
<td>Control Group</td>
<td>2.53</td>
<td>2.17</td>
<td>0.36</td>
<td>0.15</td>
</tr>
<tr>
<td>Range</td>
<td>1-4</td>
<td>1-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Range</td>
<td>1-5</td>
<td>1-3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Narrative questions. The naïve listeners were asked one question for each sample during the GIRBAS ratings (i.e., Additional observations?) and were asked narrative questions following the training session and the post-test session (Table 8).

Comments during the ratings included basic observations about the samples (e.g., “slightly instable [sic], breathiness isn't constant,” “sounds normal,” “sounds like a lot of effort to vocalize”) and reflections after reviewing the expert ratings (e.g., “Yay! I got one right,” “R[oughness] 2 & B[reathiness] 2 were different,” “I don't hear any I[nstability]--despite lack of ‘squeak,’ I still think he sounds a bit strained,” and “I still have trouble with A[sthenia] & B[reathiness].”). Narrative questions that followed the training and the post-test and summaries of subject responses follow in Table 8.
Table 8 Naïve Listener Narrative Responses

<table>
<thead>
<tr>
<th>Question</th>
<th>Themes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training 1. What was hard about rating quality?</td>
<td>Identifying qualities</td>
</tr>
<tr>
<td>Training 2. What tools or ideas helped you make the ratings?</td>
<td>Using text definitions for severity ratings</td>
</tr>
<tr>
<td></td>
<td>Using the sentence more than the /a/</td>
</tr>
<tr>
<td></td>
<td>Focused on one quality at a time</td>
</tr>
<tr>
<td></td>
<td>Auditory anchors</td>
</tr>
<tr>
<td>Training 3. How did your answers compare to the experts?</td>
<td>Usually one off</td>
</tr>
<tr>
<td></td>
<td>About half were really off/asthenia way off</td>
</tr>
<tr>
<td></td>
<td>Did better with severe examples</td>
</tr>
<tr>
<td></td>
<td>Missed qualities or added qualities</td>
</tr>
<tr>
<td></td>
<td>Tended to rate 1 versus 0</td>
</tr>
<tr>
<td>Training 4. Did you notice any trends in how your answers related to the experts?</td>
<td>Got confused with different qualities</td>
</tr>
<tr>
<td></td>
<td>Either close or way off</td>
</tr>
<tr>
<td></td>
<td>ID'ed qualities, but severity ratings didn't match</td>
</tr>
<tr>
<td></td>
<td>Severity rating were more extreme than experts</td>
</tr>
<tr>
<td>Training 5. Did you imitate the voice samples during training? When rating voices? How consistently? (Imitation group only)</td>
<td>Tried, but didn't always match</td>
</tr>
<tr>
<td></td>
<td>Some (Roughness, Breathiness, Strain) were easier than others</td>
</tr>
<tr>
<td></td>
<td>Not willingly</td>
</tr>
<tr>
<td></td>
<td>Yes, all the time</td>
</tr>
<tr>
<td>Post-test 1. Compliance rating for listening/imitation (0-100%)</td>
<td>Mean 90%, Range 75%-100%</td>
</tr>
<tr>
<td>Post-test 1b. Reason for the rating:</td>
<td>Did well</td>
</tr>
<tr>
<td></td>
<td>Misunderstood directions</td>
</tr>
<tr>
<td></td>
<td>Inattention; lack of enthusiasm</td>
</tr>
<tr>
<td>Post-test 2. What did you learn about evaluating voices during this research project?</td>
<td>Components of voice</td>
</tr>
<tr>
<td></td>
<td>Normal vs. pathological voice quality</td>
</tr>
<tr>
<td></td>
<td>The process is involved/hard</td>
</tr>
<tr>
<td></td>
<td>Imitating helped</td>
</tr>
<tr>
<td></td>
<td>Observations about own voice</td>
</tr>
<tr>
<td>Post-test 3. What was helpful to your learning?</td>
<td>Auditory anchors</td>
</tr>
<tr>
<td></td>
<td>Homework (imitating privately)</td>
</tr>
<tr>
<td></td>
<td>Feedback</td>
</tr>
<tr>
<td>Post-test 4. What would you change about the training process?</td>
<td>More examples/anchors</td>
</tr>
<tr>
<td></td>
<td>Nothing</td>
</tr>
<tr>
<td></td>
<td>Cause of disorder</td>
</tr>
<tr>
<td></td>
<td>See more expert ratings</td>
</tr>
<tr>
<td>Post-test 5. Were the anchor samples helpful? If not, what would you change?</td>
<td>Helpful (7/7 respondents)</td>
</tr>
<tr>
<td></td>
<td>Make sure the sentences all match</td>
</tr>
<tr>
<td>Post-test 6. Was it easy for you to imitate the voice samples and were some samples easier to imitate than others? Please explain or provide examples. (Imitation group only)</td>
<td>Struggled with some</td>
</tr>
<tr>
<td></td>
<td>Practice helped with accuracy</td>
</tr>
<tr>
<td></td>
<td>Increased awareness of voice sample</td>
</tr>
</tbody>
</table>
Discussion

This study investigated the influence of imitation on the perceptual abilities of naïve listeners. Results on the reliability of the expert ratings and the influence of training and imitation on listener accuracy and confidence are discussed below. Then, limitations and future directions are addressed, followed by final conclusions.

Expert Ratings

**Reliability of expert ratings.** Experts rated 25 voice samples to develop summary expert ratings and to identify anchor samples. The high general agreement (two to three experts agreeing) supports previous studies that have found high reliability between raters (Dejonckere et al., 1996; Karnell et al., 2007; Webb et al., 2004), but the low exact agreement between all three experts confirmed studies that found lower reliability (De Bodt, 1997; Yamaguchi et al., 2003). The three expert raters had high intrarater reliability, which was consistent with Dejonckere et al. (1996) and differs the from Sellars et al. (2009) finding that only one expert in their study was consistent.

**Influence of Training**

**Influence of training on accuracy.** Data confirmed that training for naïve listeners did improve the perceptual abilities and was retained at least seven days after the initial training. Numerous studies (Awan & Lawson, 2009; Chan et al., 2012; Eadie & Baylor, 2006; Ghio et al., 2014; Iwarsson & Reinholt Petersen, 2012) found that training affects the immediate accuracy of perceptual ratings, and this study confirmed those findings though the improvement varied for different voice qualities. Chan et al. (2012) found that summary feedback as compared to immediate feedback was more effective in retaining new perceptual skills five days later. Our study used delayed feedback for the training session and no feedback for the homework. We also found that post-test scores maintained improvement after the initial training for at least seven days. It is interesting that while the average ratings were the most accurate for the training session, the standard deviation was the smallest for the post-test, which suggests the scores
were more consistent across the subjects after completing a homework with no feedback and a break of at least seven days.

**Influence of training on confidence.** This study established that training improved the confidence levels of the participants in their ratings. The post-test confidence levels for the naïve listeners exceeded the confidence levels of the experts who rated the same samples. However, the high level of confidence of the participants did not correlate with expert-level accuracy. The listeners did increase accuracy from pretest to post-test, but still had a notable divergence from the expert rating. These high confidence levels may reflect familiarity and comfort with the perceptual rating scale and listening to voices, rather than a confidence level in the accuracy of the ratings themselves. However, the absence of very low confidence ratings (4 and 5) and the relative accuracy of the post-test ratings did support the idea that listeners’ confidence levels may have changed as a reflection of their perceptual skills.

**Influence of Imitation**

**Influence of imitation on accuracy.** In contrast to our initial hypothesis, the pilot study results suggested that perceptual training with imitation did not improve accuracy compared to perceptual training without imitation. Collecting additional data (N = 22) would determine if the finding that imitation did not improve performance would be found to be statistically significant.

The dual-stream theory of speech perception (Hickok & Poeppel, 2000) and the linked dual representation model of voice perception (Hutchins & Moreno, 2013) both predict that motor production experience has an influence on perception. Yet, these are theoretical models and much is still unknown about the neurological processes connecting production and perception. The stimulus provided by imitation may not intersect with the neurological connections between these two systems or the models may not apply to voice quality. The dual-stream theory has been evaluated for speech and the linked dual representation model for pitch perception, but even though speech and pitch perception are closely related to voice quality, they are not necessarily all processed in the same manner. These data do not support
expanding the linked dual representation model of voice perception from pitch perception to all aspects of voice.

It is also possible that the perception-production link is not robust enough to result in perceptual changes. The linked dual representation model of voice perception has a unidirectional link from perception to production. If it is correct that there is a unidirectional link, the feedback from imitation would be rerouted through the auditory system before it could influence perception. That longer route, versus the direct bidirectional connection of the dual-stream theory, would likely reduce the efficacy of imitation influencing perception. However, biological systems frequently have bidirectional connections rather than unidirectional connections, and neurological systems are frequently interconnected in multiple ways. It, therefore, seems more likely that a perception-production connection for voice would be bidirectional, especially if the connection for speech is bidirectional.

Beyond the theories, it is possible that imitation simply does not influence perception. Iwarsson and Reinholt Petersen (2012) included imitation exercises in a successful consensus training protocol that was embedded in a graduate voice course. However, that study did not separate out the influence of the motor experience from the other training components. Imitation may not have been a key feature in the success of that training protocol. But, there are other possible considerations to explain these results. The first consideration is whether there was enough imitation practice in this training protocol. Implicit learning is a slow process that requires many repetitions, and it is possible that this training protocol did not reach the critical threshold. DeBoer and Shealy (1995) found that 7 weeks of singing study with regular practicing resulted in perceptual changes for student subjects. This study was 3 to 4 weeks long and participants were not prompted to practice imitation on their own. Therefore, the participants had significantly less imitation experience than participants in the DeBoer and Shealy study.

It is also possible that the critical listening completed by the control group provided sufficient motor neuron stimulation to have a similar training effect as imitation. An fMRI study of finger movement versus mental rehearsal found that both practices improved performance, but used different neural pathways (Nyberg, Eriksson, Larsson, & Marklund, 2006). If critical listening provided a training
effect, then there was no true control group and the equivalent results might be expected. Also, the composition of the subject groups was not balanced. In the control group, the proportion of CSD majors and CSD seniors was higher, so participants had a larger field specific knowledge base and possible experience with other perceptual evaluations. Furthermore, the control group had significantly more instrumental music experience, which influenced perceptual skills (Kleber et al., 2010), and had higher voice self-ratings, which influenced how listeners rated other voices (Haskell, 1987). Due to these factors, the control group may have received equivalent benefit from the didactic training alone than the less experienced subjects received from the training supplemented with imitation.

Another possibility would be whether imitating a voice with organic pathologies using a healthy voice would result in relevant changes to the participants’ perceptual systems. The participants would not exactly imitate the motor production of the pathological voice because the anatomies differ. Furthermore, when healthy listeners imitate functional voice disorders, it is not known if they will use the same muscular processes and if this would make a difference in their perceptions. If the motor patterns practiced during imitation did not match the motor production of the pathological voice sample, a change in the production-perception system would not result in more accurate judgments. Finally, many of the studies evaluating reliability and/or training addressed a limited number of voice qualities. For example, Chan and Yiu (2006) only evaluated training for breathiness, and listeners completed 114–144 ratings on one quality. In our study, six different qualities were addressed and listeners completed 144 ratings, but only 24 ratings per quality. To notice a training effect with imitation, it might be necessary to focus on a smaller number of voice qualities and/or provide significantly more ratings practice.

Influence of imitation on confidence. The confidence scores for the motor and control groups both improved from pretest to post-test, with a larger gain for the imitation group. Motor practice appeared to increase confidence in the ratings but did not change the accuracy level. If this finding is established as statistically significant or is replicated, it could have important ramifications for training protocols. Increasing rater confidence beyond the rater’s skills is not an ideal training outcome.
While speech-language pathology instructors do not typically want students to be more confident than they are accurate, confidence gained from imitation may serve a purpose. First of all, the imitation group may be building the groundwork for internal standards that will eventually lead to improved accuracy. This idea correlates with anecdotal reports of clinicians using imitation and “empathic” listening to evaluate and understand clients’ vocal productions. Experience and exposure may be necessary for any benefit to be observed. Secondly, feeling more confident may encourage those students to pursue further experiences in evaluating and treating voice disorders, and those continued experiences may lead to developing more accurate judgment skills. Therefore, increased confidence without improved accuracy could still be beneficial.

However, it is possible that the change in confidence levels for the imitation group related more to group composition. The imitation group was older, with a wide standard deviation, and had lower vocal self-rating scores. One speculation is that the more experienced perspective and low vocal self-assessment (Haskell, 1987) may have led them to be more conservative in the initial confidence ratings or to be more confident in the changes in their skills from the pretest to the post-test.

**Naïve Listeners’ Narrative Responses**

Naïve listeners answered narrative questions after the training and final sessions. The challenges they identified in rating voice quality were consistent with previous studies. Rating voices was difficult because of the subjectivity, ambiguity of terminology, discriminating between severity levels and discriminating specific qualities (Kreiman & Gerratt, 2010). The participants felt they were helped by the text and auditory anchors, which is consistent with previous research into anchors (Awan & Lawson, 2009; Chan & Yiu, 2002). At least one participant felt the connected speech sample was more helpful than the sustained vowel. Evaluating speech reflects typical clinical application of the rating scale, but less research has addressed rating connected speech versus sustained vowels (Chan & Yiu, 2006; Lu & Matteson, 2014).

Participants in the imitation group were asked about their use of imitation. Subjects commented that some qualities were easier to imitate, and one participant mentioned a great reluctance to do the
imitation. It makes sense that not all qualities would be imitated as easily as others. The atypical motor movements in instability would be difficult to replicate with a healthy neurological system. Asthenia was hard for some participants to conceptualize, which is consistent with American student listeners (Yamaguchi et al., 2003), and it was thus harder to imitate. The reluctant participant required prompting and support to do the imitations, did not always do all three repetitions, and appeared less concerned about matching the voice qualities. If imitation was included in a voice disorders course, there would likely be some reluctant imitators as well as compliant imitators. It would be interesting to see how compliance and imitation accuracy influence ratings.

After the post-test, questions addressed protocol compliance, what was learned, what was helpful, and what the subjects might change about the protocol. Self-reported compliance for listening to or imitating samples three times was generally high, which supports that the improved accuracy resulted from the training. The subjects identified that they learned about the voice, about the process of evaluating voices, and about their own voices. One participant volunteered that she would now change her personal voice quality assessment from the first session. Participants found that the auditory anchors, the homework, and reviewing expert ratings helped learning. These observations are consistent with previous research on listener accuracy (Awan & Lawson, 2009; Chan et al., 2012). Participants indicated that they would like to hear more auditory examples and anchors, to know the etiology and diagnosis for the samples and to see more expert ratings. The course-based training protocol developed by Iwarsson and Reinholt Petersen (2012) included all of these components.

Following the post-test, the imitation group reiterated that they struggled to imitate some voice qualities, that practicing improved imitation accuracy, and that imitation increased their awareness of the voice qualities. These written responses were consistent with verbal comments made to the PI in conversation. Members of the imitation group emphasized that they felt imitation focused their attention and improved their ability to analyze the voice samples. The reported change in perception corroborates changes observed by students who completed singing voice study (DeBoer & Shealy, 1995), but participants in that study made actual perceptual gains with motor training. The post-test reports on the
role of imitation correlated with the increased confidence levels reported for the post-test, but the imitation group’s reported belief that imitation was helpful was not substantiated by the data.

**Limitations**

Several limiting factors should be taken into account for this study. The sample size of seven is too small to establish statistical significance. The divergence scores were highly variable between the naïve listeners and within individual listeners; therefore, trends could not be established for this small sample. When analyzed individually, some ratings for individual voice samples from pretest to post-test approached statistical significance, but more participants are needed to make broader conclusions. Based on the assumptions specified in the methodology, 22 participants (11 in the control group and 11 in the imitation group) would be necessary. It is also true that the average initial scores of the naïve listeners were relatively close to the expert rating, which is consistent with a study by Ghio et al. (2014). Given these initial scores, the ceiling effect and regression to the mean are both a concern when comparing initial scores to the post-test. Additionally, a comparison of naïve listener divergence scores and expert divergence scores would provide a perspective on typical variations in perceptual ratings.

The time frame for this study was limited (4 sessions over a 3- to 4-week period) in order to recruit participants. This was a short time for implicit and explicit learning to occur. It is possible that longer training might provide more time to solidify and stabilize learning. This may explain why at this time no difference was noted in the accuracy between the control and imitation groups in the post-test. Another factor is the severity of the selected voice samples. The voice samples were predominantly severe because selection criteria focused on samples that clearly presented different voice qualities. The large number of severe voice samples, compared to normal and mild and moderate samples, may have biased the ratings of the naïve listeners (Kreiman et al., 1993).

Additionally, the homework did not have anchor samples. There was a technical problem with the online homework, and the anchor samples did not play. The problem was not discovered until after several participants had completed the homework, so this was not changed for subsequent participants. There was an increased divergence spike on the homework for six of the seven participants (see Figure 7).
The necessary reliance on internal standards rather than reference samples may explain the increased divergence spike, but increase in divergence scores could also be related to the particular voice samples on the homework or the listening environment (e.g., the participant’s selected location and speakers/headset). Without anchor samples and controlled listening settings on the homework, this cannot be determined.

**Future Directions**

Future studies into the effects of motor experience on perceptual training would benefit from investigation into the amount of training, the role of practice and the use of anchors for qualities. Chan and Yiu (2006) noted positive training effects when the same quality was rated in 114–144 voice samples. It would be valuable to investigate the number of practice ratings and the amount of training necessary to produce perceptual changes. It would also be interesting to compare the accuracy of perceptual voice evaluations by practicing clinicians that report using imitation when evaluating voices and a comparable group of clinicians that do not regularly use imitation. That could provide more information regarding anecdotal evidence about imitation in the clinical setting and whether experience plays a part. Questions also remain about the influence of motor practice. Changes were not noted in the imitation group versus the control group, and information is needed about the interactions between the motor and perceptual systems and what intervention might be necessary to make a change.

Another area that needs further research is the use of anchors for qualities. Past research (Awan & Lawson, 2009; Chan et al., 2012; Chan & Yiu, 2006; Eadie & Baylor, 2006; Kreiman et al., 2007) used anchors for the severity levels (0–3) of one or two qualities, but this study used single anchors for each quality. It is recognized that the voice is multidimensional and that voice qualities do not appear in isolation (Kreiman et al., 1994). More information about the best practices for using voice quality anchors would be valuable and could be incorporated in the full study as discussed below.

The basic organization and structure of this pilot study was successful for addressing the hypotheses about training and imitation; however, there are aspects that could be changed if the study is continued. Small changes would include designing a randomization procedure to create more equivalent
subject groups and addressing technical issues (i.e., functional anchor samples for the homework, unintentionally advancing to the next PowerPoint slide). The pretest/post-test could be expanded to include repeated voice samples to evaluate intrarater agreement before and after training. Additionally, the use of PowerPoint for the sessions was successful for this small group of participants. With a larger study group, it would be more efficient to use a computer program that directly records participant ratings and comments in a database or spreadsheet.

Data would be collected with a least 11 subjects in each group to collect statistically significant data. Ideally, the study would be embedded in a voice disorders course and imitation experiences provided to an experimental group and control group in an alternating treatment design across an entire semester. In hindsight, it might have been more valuable and successful to draw participants for the pilot study from the graduate level instead of from undergraduates. The goal was to avoid interference from clinical training, motor speech disorders coursework and other advanced courses, but it may have been beneficial to present this training in the larger context of Introduction to Voice Disorders. Whether the protocol was expanded or remained 3 to 4 weeks long, adding regular imitation practice and encouraging independent imitation practice for the imitation group would increase the motor system demands during the study. Adding additional training examples and voice samples to rate would also enhance the training.

Expanding the pretest format and changing the process for selecting anchor samples are opportunities for improvement. The addition of a pretest without anchor samples followed by a pretest with anchor samples would provide data within the study about the benefit of anchor samples. When the experts individually suggested anchor samples, they selected many different samples and the lack of overlap between those samples meant that two of the anchors were only suggested by a single expert. It would be beneficial to have the experts meet and select the anchor samples by consensus. The experts would also have the opportunity to discuss and determine the importance of including anchor samples with commonly co-occurring qualities (e.g., breathiness-strain and breathiness-asthenia) and/or providing descriptions of the various qualities in the anchor sample. This would ensure that the anchor samples
were valid reflections of the experts’ experience and perspectives as well as valid representations of the multidimensional nature of the voice. Data could then be collected on effective practices for voice quality anchors. Furthermore, some naïve listeners critiqued the use of anchor samples with different sentences than the test samples (e.g., the strain anchor sample sentence was, “Ease off on Amy and Alan,” and most test samples were, “We were away a year ago.”) However, the more complicated task of making generalizations in order to compare different sentences may have increased the germane cognitive load and, therefore, learning (van Merriënboer, Schuurman, de Croock, & Paas, 2002). This aspect likely increased student learning and may not need to be changed.

**Conclusions**

To summarize, this study confirmed that training of naïve listeners improved perceptual evaluation skills and that these gains were retained at least seven days later. Participants also had more confidence in their ratings at the post-test. In contrast to our original hypothesis, motor training via imitation over a 4 week period did not improve accuracy when compared to the control group. The post-test scores between the imitation and control groups were equivalent. Unexpectedly, the motor training did improve the confidence levels when compared to the control group even though accuracy was not improved. However, limited sample size restricts the generalizations that can be made from these data, and another study with more participants could result in more conclusive findings.
References


