Experience and Perception: How Experience Affects Perception of Naturalness Change in Speakers with Dysarthria

THESIS

Presented in Partial Fulfillment of the Requirements for the Degree Master of Arts in the Graduate School of The Ohio State University

By

Katherine Megan Noel Kubitskey

Graduate Program in Speech-Language Pathology

The Ohio State University

2015

Master's Examination Committee:

Rebecca McCauley, Advisor

Stacy Harnish
Copyrighted by

Katherine Megan Noel Kubitskey

2015
Abstract

Introduction: This study examines the effects of listener training on naturalness ratings for subjects with Acquired Childhood Dysarthria (ACD) following acquired brain injury who had participated in a multiple case study design. Pre-treatment, post-treatment, and follow-up assessments of naturalness were obtained from two subjects with dysarthria. Samples of their speech from each subject were presented to untrained listeners, graduate students in speech-language pathology, and trained clinicians to determine whether each group would rate the speakers' naturalness differently.

Methods and Procedures: Samples of dysarthric speech from pre-treatment, post-treatment, and follow-up sessions were organized into paired sentences. 71 listeners - 29 untrained listeners, 20 graduate students, and 22 trained clinicians - assessed the pairs by determining which sentence in each pair was more natural. Assessments of each group were performed via chi-square analysis and Fisher Exact tests to determine whether there was a difference in ratings between groups.

Outcome and Results: Listener groups differed in their naturalness ratings, with untrained listeners more likely to find post treatment and follow-up sentences more natural across all listening conditions and trained clinicians the least likely to do so. Graduate students' ratings fell between the other groups. A within-group reliability analysis indicated that untrained listeners were most reliable across all ratings and trained the least reliable by a very small margin.
Discussion: These results suggest a need for further research on the effects of training on perceptual ratings. The difference in ratings calls into question whose perception of naturalness is most valuable during treatment. The inverse correlation between training and within-group reliability also bears further investigation.
Acknowledgments

I would like especially to give thanks to my advisor, Dr. Rebecca McCauley, for helping me to discover my research focus, encouraging me to keep going, and pushing me to do my best during this process.

To Jennifer Lundine who allowed me to join her investigation of dysarthria the effects of Lee Silverman Voice Therapy on acquired childhood dysarthria and provided valuable recommendations regarding presentation of my work.

To Dr.s Harnish and Ellawadi and Michael Bringold I wish to express my gratitude for their support in analyzing the data that made up my conclusions; and to Laura Wadsworth, who helped organize the data.

To the speakers and the listeners who contributed to my research, I appreciate your time and willingness to contribute to the evidence that keeps the field of speech-language pathology strong.
Vita

May 2006 ..............................................West Anchorage High School

2010........................................................B.A. English, U. of Alaska Anchorage

2014........................................................M.A. English, U. of Alaska Anchorage

2013 to present .................................M.A. Speech Language Pathology, The Ohio State University

Fields of Study

Major Field: Speech-Language Pathology

   English: Rhetoric and Language
Table of Contents

Abstract ................................................................................................................................. ii

Acknowledgments .............................................................................................................. iv

Vita ....................................................................................................................................... v

List of Tables ....................................................................................................................... vii

Chapter 1: Literature Review ............................................................................................ 1

Chapter 2: Methods ............................................................................................................. 11

Chapter 3: Results ............................................................................................................... 16

Chapter 4: Discussion ......................................................................................................... 21

References ........................................................................................................................... 24

Appendix A: Demographic Questionnaire ......................................................................... 28

Appendix B: Procedural Script ......................................................................................... 29

Appendix C: Ratings Forms ............................................................................................. 30
List of Tables

Table 1. PRE vs. POST, Speaker 1 ................................................................. 16
Table 2. PRE vs. POST, Speaker 2 .................................................................. 17
Table 3. POST vs. FU, Speaker 1 ................................................................... 18
Table 4. POST vs. FU, Speaker 2 ................................................................... 18
Table 5. PRE vs. FU, Speaker 1 ................................................................... 19
Table 6. PRE vs. FU, Speaker 2 ................................................................... 19
Chapter 1: Literature Review

Introduction

Humans have developed verbal communication as a vital piece of interaction; it allows them to express and share needs and desires through a commonly accepted medium accessible to most. Neurological damage, such as that following brain injury, may impair verbal communication in ways that limit the speaker's independence. Speakers with acquired dysarthria experience a change in their speaking pattern that often leads to reduced intelligibility and/or naturalness. Interventions for dysarthria tend to focus first on over-articulation and reduced rate to improve intelligibility; these strategies often result in increased intelligibility but decreased naturalness (Dagenais et al., 2006; Duffy, 1995; Logan, Roberts, Pretto, & Morey, 2002; Yorkston, Beukelman, Strand, & Bell, 1999). This effect is less pronounced with more severely dysarthric speakers, presumably due to the fact that the speakers' naturalness baseline is already so low that the strategies did not create a further detrimental effect (Yorkston et al., 1990). However, less dysarthric speakers have shown a marked decrease in naturalness, which may limit their speaking opportunities (Dagenais, 2006).

Listeners form judgments based on a speaker's vocabulary, speech range, pitch, and accent. These judgments may be positive or negative, conscious or subconscious, and may lead to a change in the way the speaker is viewed and treated by the listener.
Dagenais et al.'s designation of naturalness as "acceptability" (1998, 1999, 2006) points to the very problem of allowing naturalness to decrease in favor of other speech variables. As listeners come to view speech as less acceptable, they will be less inclined to listen to the speaker and therefore place limits on the speaker's apparent communicative abilities.

In this study, I look at the effect of listeners' level of clinical experience on their perception of the naturalness of treated vs. untreated dysarthric speech. The speakers and their family were asked in a previous study (Lundine & Harnish, 2014) to perceptually rate their own progress toward improved speech at pre-treatment, post-treatment, and follow-up sessions. The ratings were based on a number of variables commonly associated with naturalness. The speakers rated their progress differently than their parents and the clinician, raising the question of whose ratings were most important in judging their change in speech quality. The difference in ratings is in line with previous research indicating that listener groups who vary in their overall experience with pathological speech also vary in their rating of dysarthric speech (Kreiman et al., 1990). The current study examines the effect of experience by enlisting untrained, graduate-student, and trained listeners to rate the naturalness of dysarthric speech recorded at pre-treatment, post-treatment, and follow-up sessions.

Dysarthria refers to a class of disorders in which movements that allow for clear speech are disrupted due to muscular weakness or incoordination. The class is composed of several subtypes defined by the analysis of 38 speech features (Darley et al., 1969; Duffy, 1995; Dykstra, 2007) that are affected by a variety of neurological conditions. It is
difficult to quantify the prevalence or subtypes of dysarthria among children due to different methods used to measure dysarthria, assessments performed at different stages post-injury, and the variety of populations assessed (van Mourik et al., 1997; Murdoch et al., 2001; McDonald et al., 2013). Despite the variation in reporting, most sources agree that dysarthria is a persistent disorder that significantly alters a person's ability to speak, often limiting their socialization and occupational opportunities (Dykstra, 2007). This is due in part to decreased naturalness of the speaker's voice, which negatively affects listeners' perceptions of the speaker.

Parameters of dysarthric speech that may be distorted include range, timing, speed, and steadiness of speech mechanism movements (Bellaire et al., 1986). These variables are often correlated with perceived distortions in naturalness and intelligibility. When naturalness is judged to be poor, intelligibility may also suffer. Viewing these deficits through the lens of the WHO ICF is valuable to show their effect on the impressions of inexperienced listeners (Brunnegard et al., 2009; Dykstra et al., 2007; Lowit & Kent 2010; World Health Organization, 2002). The presence of dysarthria does not necessarily indicate a limiting disability; however when the characteristics of dysarthria reduce individuals' speech capacity and performance to the point of affecting the perceptions of listeners around them, the disorder restricts their communicative participation and keeps them from optimal performance in functional living situations, thereby becoming a disability. As a result, naturalness is an important goal for speech therapy as a part of improving speaker's ability to communicate.
**Naturalness**

As a goal, increased naturalness must have a measurable outcome in order to justify its place in evidence-based practice (EBP). EBP refers to the process by which speech language pathologists evaluate treatment outcomes for effectiveness and choose appropriate, theory-driven treatments for their own clients. Using EBP increases the likelihood that clinicians can provide optimal treatment and achieve fidelity in treatment outcomes for multiple clients (Kaderavek & Justice, 2010). Little research exists regarding the reliability and validity of naturalness (also referred to as acceptability or bizarreness) measures employed in current therapeutic interventions for assessing pre- and post-intervention naturalness. When researching this characteristic, researchers vary in their designs for ratings forms. Witt et al. (1997) described the use of three separate rating scales for speech acceptability - yes/no questionnaires asking naive listeners about the occurrence of 12 vocal qualities (including stuffy nose, difficult to understand, loudness, and speed of speech); yes/no questions for the speakers asking them whether they wish their voices were different; and a Likert 5-point equal interval scale for clinicians judging 5 vocal qualities (pitch, intensity, resonance, rate, intelligibility). The scales were developed by "team professionals" whose credentials and process for development were not disclosed. Tamplin et al. (2008) also presented different rating scales for each group, designed by trained speech pathologists whose credentials were also undisclosed. The use of separate scales for each group in both studies prevented a direct comparison of judgments between groups due to the potential effect of different
descriptors for each group as opposed to their differing levels of experience on their ratings.

Mahler and Ramig (2012) asked listeners to rate speakers as compared to each other using a 100 point scale, simultaneously evaluating loudness, voice quality, overall pitch variability and articulation clarity. Kreiman et al. instructed their listeners - two groups of 5 listeners representing expert clinicians and naive listeners - to complete 7-point scales asking how different each sentence pair was. Dagenais et al. (2006) asked normal adult listeners of unknown experience to rate sentences via a 9-point equal interval scale ranging from "terrible" through "excellent". None of the studies described above used measures of naturalness with standardized norms and all of the measures differed in choice of characteristics rated. This illustrates the difficulty clinicians face in finding a standardized, widely used, evidence-based measurement of naturalness; one result of such different ratings is that it is difficult to compare ratings across studies and across groups of listeners to determine the most salient elements for perception between groups.

Some factors that have been hypothesized to affect naturalness include speakers' intelligibility, vocal intensity or stress, pitch range, rate of speech or use of pauses, articulatory precision and prosody (Tamplin et al., 2008; Walshe et al., 2008). Moreover, the listeners' own experience - training, attitudes, and/or preconceived judgments - have also been shown to influence their perception and subsequent rating of naturalness for a speaker (Dagenais et al., 1998, 1999; Brunnegard, 2009). As a result, clinicians design
interventions based on the particular speech factors they judge to be most important in changing naturalness for dysarthric speakers.

In designing and choosing interventions for speakers with dysarthria, clinicians work to improve communicative function by focusing on specific speech subsystems, such as the velopharyngeal system for improved nasality, nasal emission, and intelligibility (Yorkston et al., 2001; Yorkston et al., 2007). Others attempt speech system change on a global level, choosing speech characteristics designed to alter the speaker's entire utterance at once. Interventions focusing on global change have used loudness, speaking rate, and prosody as agents of change. Of those factors, loudness has been best documented to show effective system-wide change in speakers with dysarthria (Fox et al., 2012; Yorkston et al., 2007). Specifically, Lee Silverman Voice Therapy (LSVT) is the most well-described, precise protocol that has yet been linked to physiologic and perceptual features of dysarthria. Positive changes in loudness and other aspects of speech production have been documented in LSVT outcomes (Fox & Boliek, 2012; Fox et al., 2012; Mahler & Ramig, 2012; Sapir et al., 2003; Wenke et al., 2008); there remains, however, limited information on naturalness outcomes following LSVT.

**Value of naturalness for LSVT treatment in individuals**

LSVT is designed to increase an individual’s loudness, a factor that has been linked to both intelligibility and naturalness (Fox et al., 2012; Tamplin et al., 2008; Walshe et al., 2008). It was originally developed for individuals with Parkinson's disease (PD), whose comorbid, progressive speech and voice disorders can result in a variety of symptoms including in soft voice, monotone, breathiness, hoarse voice quality, and
imprecise articulation, negatively impacting their ability to communicate. Previous interventions did not provide long-term, significant changes to their speech, partially due to uncertainty as to which mechanisms are causing the changes and therefore should be addressed (Fox et al., 2012). LSVT has since been assessed (Ramig et al., 1995, 1996, 2001a, 2001b) and designated as an effective, evidence-based intervention to improve the communication abilities of individuals with PD and other disorders that lead to similar speech outcomes (Fox & Boliek, 2012; Fox et al., 2012; Sapir et al., 2003; Yorkston et al., 2007).

One advantage of LSVT is that it has one explicit goal: to speak more loudly. The single goal may reduce the cognitive load on patients whose cognitive functions have declined as a result of dementia, a common concomitant disorder for patients with Parkinson's disease. The low cognitive load may also be beneficial for patients who have suffered a traumatic brain injury or other trauma to the brain. Speaking more loudly has been shown to induce system-wide improvements in the speech of individuals with PD, MS, and hypokinetic dysarthria, creating changes in overall intelligibility and articulatory precision following the achievement of a louder voice (Fox & Boliek, 2012; Ramig, Sapir, Countryman, et al., 2001; Ramig, Sapir, Fox, & Countryman, 2001; Sapir et al., 2001; Sapir et al., 2003; Wenke et al., 2008). These results suggest that further investigation may be useful in determining whether LSVT's effects extend to inducing change in naturalness as well.
Perceptual ratings and experience

Previous studies have examined listeners' perceptions of pathological voices, assessing how listeners rate disordered voices and whether different groups of listeners rate voices similarly. Groups included untrained listeners, graduate students, and trained clinicians; pathologies included stuttering, dysarthria, voice disorders, and alaryngeal speech (Dagenais et al., 1999, 2006; Finizia et al., 1998; Martin et al., 1984; O'Brian et al., 2003; Van As et al., 2003). In a study of voice quality ratings, Kreiman et al. (1990) found that listeners rely on different voice parameters depending on their level of experience. When asked to judge whether two disordered voices were similar, naive listeners were more likely to rely on the differences in fundamental frequency, whereas expert listeners paid more attention to breathiness and roughness of the voices. The naive listeners' resulting judgments of the voices were more homogenous between speaker pairs than those of expert listeners. In a series of studies by Dagenais and his colleagues (Dagenais et al., 1998; Dagenais et al., 1999; Dagenais and Wilson, 2002), when asked to rate the naturalness of speakers with mild and moderate dysarthria, trained listeners rated speakers significantly lower than untrained listeners for speakers with moderate dysarthria; ratings across listener groups were fairly homogenous for speakers with mild dysarthria. This finding suggests that the perceptions of the professional do not always match those of the community regarding the naturalness of speakers with dysarthria. However, it is not clear exactly what effect experience has on listener judgment. Results from these studies have indicated that trained and untrained listeners apply different processes to their judgments of voice quality; it has been hypothesized that trained
listeners incorporate their experience of pathological voices, leading to somewhat different decisions regarding voice quality than their untrained counterparts.

**Rationale for this study and research questions**

This study continues the investigation of differences across listener groups' judgments of naturalness of dysarthric speech. Dysarthric speech becomes a disability when it impairs an individual's ability to communicate and participate in functional living activities. Though clinician judgment of speech is a crucial factor in deciding whether an individual requires services, it is important to know how closely clinician judgment resembles that of the untrained listeners who surround the individual every day given that it is the latter group of listeners who comprise the more typical listener group for individuals with dysarthria.

Lundine and Harnish (2014) conducted a study of two cases of childhood acquired dysarthria to determine whether LSVT has an effect on acoustic and perceptual parameters of speech and communication: loudness, shaky voice, monotone, strain, mumbling, and intelligibility. The current study used the speaker recordings from the Lundine and Harnish study as a basis for perceptual ratings from listeners of different levels of experience with pathological voices to examine how that experience may contribute to listener groups' differing judgments of speaker naturalness. The recordings captured the speakers' speech before (PRE), immediately following (POST), and 3 months following (FU) participation in LSVT protocol. For each speaker, sentences from the recordings were randomly paired and presented to untrained, graduate student, and trained listeners. The listeners blindly rated speaker naturalness before, during and after
LSVT protocol in an effort to answer the following research questions: for two young speakers who underwent LSVT to treat their dysarthria, do listeners' level of experience affect their perception of speaker naturalness? Are there differences in reliability between groups? Are there differences in listener agreement within groups?
Chapter 2: Methods

Speech Stimuli

The speech stimuli used in this study were produced by two participants from an associated treatment study (Lundine et al., 2014). The two speakers with acquired childhood dysarthria (ACD) had been selected on the basis of age (under 21), persistence of dysarthria, and time following injury (six or more months post-injury). One speaker, age 19 (hereafter referred to as Speaker 01), acquired flaccid dysarthria secondary to a traumatic brain injury. The other speaker, age 13 (hereafter referred to as Speaker 02), acquired cerebellar mutism following a posterior fossa tumor resection. His cerebellar mutism later resolved into mixed flaccid-spastic-ataxic dysarthria, as diagnosed by a certified speech-language pathologist with 13 years of experience with neurogenic communication disorders. Both speakers entered Lee Silverman Voice Therapy (LSVT), a program of treatment that included goals related to changing speech rate and improving intelligibility.

Within the Lundine and Kubitskey (2015) study, the speakers were recorded producing stimuli from the Assessment of Intelligibility of Dysarthric Speech (AssIDS) (Beukelman & Yorkston, 1981; Butler, 1989; Carpenter, 1989). The AssIDS is a standard protocol designed to help clinicians quantify intelligibility and speaking rate in adolescent and adult speakers with dysarthria. It is composed of sentences of varying
length and has been widely accepted as an effective tool for determining intelligibility in speakers with dysarthria (Dagenais et al., 2006; Sapir et al., 2003; Walshe, 2008; Wenke et al., 2008). It was used so that the same stimuli were produced by both participants in the treatment study. The stimuli used for this study were of recorded sentences from the AssIDS recorded at pre, post, and follow-up sessions.

**Stimulus Recordings and Listener Judgment Files**

As part of Lundine and Kubitskey (2015), stimuli were recorded in a small treatment room using an Olympus WS-801 audio recorder and Go-Mic-by-Samson portable usb microphone. The microphone was mounted on a laptop; distance from the subject's mouth to the microphone was measured several times throughout each treatment session to maintain the target distance of 50 centimeters. The sentences were recorded at the beginning of each session, creating one file for each set of sentences. Each speaker was instructed to read a set of 10 sentences from the AssIDS stimuli; the sentences were between 10 and 14 words in length. Recordings from 6 of the 10 sentences at different points in time were chosen at random to comprise the stimuli in the current study.

Stimulus sentences for each participant were concatenated into one audio file that was then placed on a secured thumb drive for presentation to the listeners. Within each participant's recordings, sentence pairs comprised of the same sentence produced at two points in treatment (pre-post, pre-follow up, or post-follow up) were created using Adobe Audition editing software; the pairs were then sequenced in a random order. Intervals between sentences were 2 seconds long; following a five-second silent interval, each sentence pair was replayed prior to listeners' recording of their preference.
Participants

Seventy-one listeners were recruited as participants for this study following its approval by the Institutional Review Board. Before participating, they were screened to ensure that their hearing was within normal limits (see procedures below). They were offered no payment to participate in the procedure beyond having their parking validated and they were informed that they could withdraw from participation at any time. All participants signed informed consents prior to participation. They fell into three groups based on experience with dysarthric speakers: trained listeners who were experienced speech-language pathologists (n = 22), graduate student listeners who were enrolled in an SLP training program (n = 20), and naïve, or untrained, listeners who had minimal to no experience with clinical listening tasks (n = 29). The recruitment and characteristics of each group is described in greater detail below.

*Experienced SLP listeners.* Twenty-two experienced clinical speech-language pathologists aged 27-42 years were recruited by extending an invitation to clinicians working at a large Midwestern university and a large urban children’s hospital. They were not aware of the purpose of the study nor were they familiar with any of the speakers who were included as participants in the original treatment (Lundine & Kubitskey, 2015).

*Graduate student listeners.* Twenty graduate student listeners were recruited from the second-year class of students pursuing a 2-year master’s program in speech-language pathology at a large Midwestern university. They were chosen for their exposure to clinical listening tasks associated with their completed clinical coursework and
preliminary experiences as clinicians. They were recruited through invitations via email and class presentation.

_Inexperienced listeners._ Twenty-nine untrained listeners were recruited from undergraduate students taking an introductory speech-language pathology course. They were expected to have no formal training on clinical listening tasks or clinical experiences with individuals with dysarthria. They were recruited from an undergraduate course taught by faculty in the speech-language pathology department. These individuals represented the everyday listener with whom the speakers would come in contact: relatives, coaches, store employees, friends, strangers. Though the students were taking courses in speech-language pathology, they were considered untrained due to their lack of treatment experience for individuals with motor speech disorders, as well as their lack of clinical exposure to speech disorders in which they were required to make careful listening judgments (Brunnegard et al., 2009). They were provided with an incentive of a small number of bonus points on their grade for participating, commensurate to a separate extra credit opportunity; however, failing to participate did not harm their course grade.

**Listener judgment procedures**

The final audio file of speaker sentences was played over the loudspeakers in one of four rooms: three similarly-sized classrooms at a large Midwestern university, and a conference room at a nearby children's hospital. Prior to each listening session, the rooms were measured to ensure ambient noise levels below 55 dBA (Kemp & Ryan, 1993). The participants had their hearing screened using pure-tone audiometers set at 1000, 2000,
and 4000dB (ASHA, 1997); all participants passed the screening and were therefore considered to having hearing within normal limits.

All groups of listeners rated the naturalness of the speech samples presented to them by choosing the one of a pair of sentences that they regarded as the more natural of the two. The two-choice preference provided for a simple decision in a familiar format for all listeners (Sapir et al., 2003). These judgments were made after hearing each sentence pair twice. Listeners re-rated 33% of the sentences pairs provided so that a measure of test-retest reliability could be obtained (Dagenais, et al. 2006). The samples were presented in a large classroom over the classroom speakers set at maximum volume. The listening task took an average of 40 minutes for each listener.

Analysis

Resulting group differences in levels and variability of ratings were analyzed using chi squares to investigate whether membership in each listener group affected listener rating of naturalness in dysarthric speech.

The ratings were divided into four conditions: preference for pre-treatment sentences (PRE); preference for post-treatment sentences (POST); preference for follow-up sentences (FU), and no preference between PRE, POST, and FU (NP). Preference for each condition was compared across listener groups.

Reliability measurements for each listener group were also analyzed to assess consistency in listener ratings for each sentence. Results are presented below.
Chapter 3: Results

71 listeners participated in this study. The listeners were divided into three categories: untrained listeners (n=29), graduate students (n=20), and trained SLPs (n=22).

Chi-squares were used to analyze listener preference across groups for each listening condition: post (POST) vs. pre-treatment (PRE), follow-up (FU) vs. PRE, and FU vs. POST. Ratings were divided into preference for POST sentences as indicated by rating POST sentences as more natural at both opportunities; preference for PRE sentences as indicated by rating PRE sentences as more natural at both opportunities; preference for FU sentences as indicated by rating FU sentences as more natural at both opportunities; or no preference (NP) as indicated by rating one of each condition as more natural. Results of the chi squares are reported below.

Post-treatment vs. pre-treatment

<table>
<thead>
<tr>
<th>Speaker 1</th>
<th>Untrained</th>
<th>Graduate</th>
<th>Trained</th>
<th>Marginal totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>POST = 2 (2 votes for POST)</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>NP = 1 (1 vote for PRE and 1 vote for POST)</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>23</td>
</tr>
<tr>
<td>PRE = 0 (2 votes for PRE)</td>
<td>23</td>
<td>17</td>
<td>16</td>
<td>46</td>
</tr>
<tr>
<td>Marginal Totals</td>
<td>29</td>
<td>20</td>
<td>22</td>
<td>71</td>
</tr>
</tbody>
</table>

Table 1: PRE vs. POST, Speaker 1
Speaker 1. Chi square results were not significant for the distribution of listeners rating PRE vs. POST sentences (p = .66). Inspection of the data suggests that each group chose each condition at a similar rate. Proportionally, graduate students were most likely to choose PRE, followed by untrained listeners and then trained listeners.

<table>
<thead>
<tr>
<th>Speaker 2</th>
<th>Untrained</th>
<th>Graduate</th>
<th>Trained</th>
<th>Marginal totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>POST = 2</td>
<td>16</td>
<td>9</td>
<td>7</td>
<td>32</td>
</tr>
<tr>
<td>NP = 1</td>
<td>9</td>
<td>10</td>
<td>9</td>
<td>28</td>
</tr>
<tr>
<td>PRE = 2</td>
<td>4</td>
<td>1</td>
<td>6</td>
<td>11</td>
</tr>
<tr>
<td>Marginal totals</td>
<td>29</td>
<td>20</td>
<td>22</td>
<td>71</td>
</tr>
</tbody>
</table>

Table 2: PRE vs. POST, Speaker 2

Speaker 2. Chi square results were not statistically significant for listeners rating PRE vs. POST for Speaker 2 (p = 0.19). Based on visual inspection of the data, the groups' ratings were widely distributed between preference for POST, preference for PRE, and NP. Although not associated with a significant association, untrained listeners appeared to show a preference for Speaker 2’s POST sentences as more natural 55% of the time, compared to 45% of the time for graduate listeners and 32% for trained listeners.
Post vs. Follow-up

<table>
<thead>
<tr>
<th>Speaker 1</th>
<th>Untrained</th>
<th>Graduate</th>
<th>Trained</th>
<th>Marginal totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>FU preferred = 2 (2 votes for FU)</td>
<td>4</td>
<td>5</td>
<td>10</td>
<td>19</td>
</tr>
<tr>
<td>NP = 1 (1 vote for FU and 1 vote for POST)</td>
<td>12</td>
<td>13</td>
<td>10</td>
<td>35</td>
</tr>
<tr>
<td>POST = 2 (2 votes for POST)</td>
<td>13</td>
<td>2</td>
<td>2</td>
<td>17</td>
</tr>
<tr>
<td>Marginal Totals</td>
<td>29</td>
<td>20</td>
<td>22</td>
<td>71</td>
</tr>
</tbody>
</table>

Table 3: POST vs. FU, Speaker 1

**Speaker 1.** Chi square results revealed that the distribution of votes for POST vs. FU were significantly related to group membership (p = .0045). Visual inspection of the data reveals that graduate listeners selected the NP condition at a higher rate than untrained listeners and that trained listeners selected the FU condition at a similar rate that untrained listeners chose POST. Further inspection suggests that trained listener ratings were evenly distributed between FU and NP, while untrained listener ratings were evenly distributed between NP and POST. Graduate students' ratings were most concentrated in NP.

<table>
<thead>
<tr>
<th>Speaker 2</th>
<th>Untrained</th>
<th>Graduate</th>
<th>Trained</th>
<th>Marginal totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>FU = 2 (2 votes for FU)</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>NP = 1 (1 vote for FU and 1 vote for POST)</td>
<td>19</td>
<td>10</td>
<td>13</td>
<td>42</td>
</tr>
<tr>
<td>POST = 2 (2 votes for POST)</td>
<td>7</td>
<td>8</td>
<td>5</td>
<td>20</td>
</tr>
<tr>
<td>Marginal totals</td>
<td>29</td>
<td>20</td>
<td>22</td>
<td>71</td>
</tr>
</tbody>
</table>

Table 4: POST vs. FU, Speaker 2

**Speaker 2.** Chi square results were not statistically significant for listeners rating POST vs. FU for Speaker 2 (p = 0.62).
Pre vs. Follow-up

<table>
<thead>
<tr>
<th>Speaker 1</th>
<th>Untrained</th>
<th>Graduate</th>
<th>Trained</th>
<th>Marginal totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>FU = 2 (2 votes for FU)</td>
<td>7</td>
<td>3</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>NP = 1 (1 vote for FU and 1 vote for PRE)</td>
<td>13</td>
<td>10</td>
<td>2</td>
<td>25</td>
</tr>
<tr>
<td>PRE = 2 (2 votes for PRE)</td>
<td>9</td>
<td>7</td>
<td>19</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>29</td>
<td>20</td>
<td>22</td>
<td>71</td>
</tr>
</tbody>
</table>

Table 5: PRE vs. FU, Speaker 1

**Speaker 1.** Chi square results revealed that the distribution of votes for PRE vs. FU were significantly related to group membership (p = .0011). Inspection of the data suggests that untrained ratings were widely varied between each condition, whereas trained listeners were concentrated in PRE. Both groups seemed to choose PRE more often than the other two conditions; graduate students seemed far more likely to choose PRE, whereas untrained listeners were evenly divided between PRE and NP.

<table>
<thead>
<tr>
<th>Speaker 2</th>
<th>Untrained</th>
<th>Graduate</th>
<th>Trained</th>
<th>Marginal totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>FU = 2 (2 votes for FU)</td>
<td>14</td>
<td>12</td>
<td>6</td>
<td>32</td>
</tr>
<tr>
<td>NP = 1 (1 vote for FU and 1 vote for PRE)</td>
<td>11</td>
<td>5</td>
<td>10</td>
<td>26</td>
</tr>
<tr>
<td>PRE = 2 (2 votes for PRE)</td>
<td>4</td>
<td>3</td>
<td>6</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>29</td>
<td>20</td>
<td>22</td>
<td>71</td>
</tr>
</tbody>
</table>

Table 6: PRE vs. FU, Speaker 2

**Speaker 2.** Chi square results were statistically significant for listeners rating PRE vs. FU for Speaker 2 (p = 0.02). Based on visual inspection of the data, untrained listener ratings appeared to be distributed most evenly between FU (48%) and NP (38%). Graduate listeners preferred FU 60% of the time, while trained listeners preferred FU 27% of the time. No group preferred PRE more than 27% of the time.
Reliability

Listeners re-rated the same percentage of sentences the same for Speaker 2 for both the PRE-POST and the PRE-FU conditions: untrained listeners showed 66% agreement; graduate listeners showed 65% agreement, and trained listeners showed 59% agreement. All listeners groups were more likely to re-rate sentence pairs the same for Speaker 1's PRE-POST condition than his PRE-FU condition by 9-24% points. Untrained listeners varied the most, with 93% agreement for Speaker 1's PRE-POST condition and 69% for his PRE-FU condition. Graduate listeners demonstrated 85% PRE-POST agreement and 70% PRE-FU agreement; trained listeners demonstrated 86% PRE-POST agreement and 77% PRE-FU agreement.

Listeners rated Speaker 1’s PRE-POST reliability sentence pairs the same 89% of the time for Speaker 1 and 65% of the time for Speaker 2. Reliability ratings for Speaker 2 were lower than for Speaker 1 across all groups, with untrained and trained listeners deviating the most with 27 percentage points difference between Speaker 1 and Speaker 2 reliability. Graduate listeners were more consistent, with 85% agreement for Speaker 1 and 65% agreement for speaker 2.

Listeners rated the PRE-FU sentence pairs the same 70% of the time for Speaker 1 and 63% of the time for Speaker 2. Reliability ratings for Speaker 2 were lower than for Speaker 1 across all groups, with trained listeners deviating the most at 18 percentage points difference between Speaker 1 and 2. Untrained listeners were more consistent, with 69% agreement for Speaker 1 and 66% agreement for Speaker 2.
Chapter 4: Discussion

This study examined the effects of experience on listener ratings of naturalness of dysarthric voices following LSVT protocol. Untrained, graduate student, and trained listeners were asked to choose the more natural sentence from pairs with the assumption that post-treatment sentences for both speakers should sound the most natural. The difference between ratings by untrained, graduate, or trained listeners were found to be significant some of the time, revealing a possible association between experience and listener judgment.

Naturalness between speakers

Chi square results revealed that there was a significant relationship between group membership and listener ratings for Speaker 1, whereas ratings for Speaker 2 were more widely distributed across conditions and less distinct between groups. Reliability measures were similar, with all groups rating Speaker 1’s sentence pairs more reliably than Speaker 2. This finding suggests that the speakers' level of dysarthria and/or their response to LSVT played a confounding part in listener rating analysis. Overall, listeners found Speaker 1’s later-condition sentences to be less natural than those of Speaker 2.
Naturalness between groups

Listener ratings varied widely between groups. The PRE vs. FU condition showed a significant relationship between group membership and ratings for both speakers, with trained listeners preferring PRE most often and untrained listeners varying most widely in their distribution of preferences between conditions. Graduate students chose NP most often for Speaker 1, and FU for Speaker 2. In the other conditions showing significance, untrained listeners varied the most widely in their distribution; graduate and trained listeners were most likely to choose PRE over POST; and trained listeners were more likely than graduate students to prefer FU over POST.

Effects of experience

Naturalness ratings indicated that untrained listeners were most reliable in their ratings for the POST vs. PRE comparison for both speakers and for Speaker 2’s FU vs. PRE comparison; they were least consistent in their preference of earlier vs. later conditions. Trained listeners were most reliable in their ratings for Speaker 1’s FU vs. PRE comparison; they were more likely to prefer earlier-condition sentences. Graduate students' reliability ratings fell between untrained and trained listeners; their ratings tended to fall in the NP condition for most comparisons.

Limitations

The small number of speakers to provide samples of dysarthric speech poses a barrier for generalization and validity of the results. Future research should include a greater number of speakers controlled for the effects of age, severity of injury, type of treatment and type of dysarthria on the results of listener ratings.
The choice to use only a structured speaking task also places limitations on its usefulness. The use of stimuli specifically designed to assess speakers with dysarthria leads to a reproducible procedure and limits the effect of individual word choice that would arise if spontaneous speech samples were used; however, the exclusion of conversational speech meant that listeners were not able to judge speakers' naturalness in conversation - the scenario in which they are most likely to be judged by untrained listeners (Lowit et al., 2001; Ingham & Riley, 1998).

A further limitation is the lack of one standard definition of naturalness among listeners. Future research should include a method to collect listeners' definitions of naturalness to assess whether different groups are measuring naturalness in the same way.

**Summary and Conclusions**

Despite the study's limitations, it appears that experience with disordered voices may relate to listeners' naturalness ratings of speakers with dysarthria. The results from this study support other findings (Dagenais et al., 2006; Tamplin, 2008; Kreiman et al., 1990) in research regarding naturalness perceptions. Future research should continue to explore exactly how the groups differ in their naturalness ratings for speakers with dysarthria.
References


Appendix A: Demographic Questionnaire

Age:
Gender:
Occupation:

*If graduate student:*

- Year in program:

- Coursework taken that relates to dysarthric speakers:

- Nature of experience with dysarthric speakers:
  - 1-5 per month
  - 6-10 per month
  - 11 or more per month
  
  - Approximate number of treatment cases involving this population:
    - 1-5 per month
    - 6-10 per month
    - 11 or more per month

*If speech language pathologist:*

- Number of years as a practicing clinician:

- Nature of experience with dysarthric speakers:

  - Approximate number of assessments involving this population:
    - 1-5 per month
    - 6-10 per month
    - 11 or more per month

  - Approximate number of treatment cases involving this population:
    - 1-5 per month
    - 6-10 per month
    - 11 or more per month
Appendix B: Procedural Script

*Hearing screening:* Please wear these headphones, the red cup on your right ear and the blue on your left ear. As you hear each tone, please raise your hand on the side you hear the tone.

*Listening experiment:* Hello. My name is <facilitator>. These are <screener names>; they will be screening your hearing before the experiment. Let me start by thanking each of you for coming. We appreciate your taking the time to come today.

Today you will be listening to 16 pairs of sentences and rating them according to the forms in front of you. If at any time you feel uncomfortable or wish to leave, you are completely free to do so without any consequences. Any questions?

<instructions for intelligibility portion, explanation of form>

You will hear the same speakers for eight more pairs of sentences. Each pair will be played once. As you listen to each pair, think about which sentence sounds more natural - sentence 1 or sentence 2. Indicate your choice on the form provided by circling the more natural sentence.
Appendix C: Rating Forms

Naturalness

Please indicate which sentence is more natural by circling "Sentence 1" or "Sentence 2"

1. Sentence 1 Sentence 2
2. Sentence 1 Sentence 2
3. Sentence 1 Sentence 2
4. Sentence 1 Sentence 2
5. Sentence 1 Sentence 2
6. Sentence 1 Sentence 2
7. Sentence 1 Sentence 2
8. Sentence 1 Sentence 2