Utility of Perceived Listener Effort as an Outcome Measure for Disordered Speech and Voice

Kathleen F. Nagle

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2014

Reading Committee:
Tanya Eadie, Chair
Kristie Spencer
Kathryn Yorkston

Program Authorized to Offer Degree:
Speech & Hearing Sciences

University of Washington

Abstract

Utility of Perceived Listener Effort
as an Outcome Measure for Disordered Speech and Voice

Kathleen F. Nagle

Chair of the Supervisory Committee:
Associate Professor Tanya Eadie
Speech & Hearing Sciences

Listening effort is a concept that has been shown to be clinically useful measuring hearing outcomes. Given the potential effects of disordered speech or voice on normal-hearing listeners, the effort required to listen to such speech may be worth evaluating for disordered speech as well. This document includes a review of the literature (Chapter 1) on objectively measured listening effort and subjectively measured *perceived listener effort* (PLE), specifically to determine the utility of PLE ratings for disordered speech and voice. Following this review, three studies are described in which PLE is compared to current outcome measures (i.e., speech acceptability and intelligibility) using tracheoesophageal (Chapter 2) and electrolaryngeal (Chapter 3) speech samples. Despite strong correlations among the measures, ratings of PLE were shown to differ significantly depending on sentence length (number of words), whereas ratings of acceptability and intelligibility did not. Qualitative data reviewed in Chapter 4 revealed similar interpretations of the terms "acceptability" and "listener effort" by everyday listeners; "understandability" was reported as a major component of both. However, listeners described differences between the concepts as well. Acceptability was interpreted as pleasantness, whereas PLE was specifically how difficult it was to listen to the speech samples. Findings support continued research into the factors affecting PLE and how best to measure it for evaluations of disordered speech and voice.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ORGANIZATION

This dissertation is organized as four self-contained manuscripts with a common theme of investigating perceived listener effort. Each manuscript was written in preparation for publication and can be read independently of the others. However, there is overlap in background information and references cited among the chapters, particularly Chapters 2 and 3.

Chapter 1: Nagle, K.F. "Objective versus subjective methods of measuring listening effort: A literature review," to be submitted for publication.

Chapter 2: "Perceived listener effort for highly intelligible tracheoesophageal speech" published with minor changes as Nagle, K. F & Eadie, T.L. (2012). Listener effort for highly intelligible tracheoesophageal speech. *Journal of Communication Disorders, 45,* 235-245.

Chapter 3: Nagle, K.F. & Eadie, T.L. "Utility of perceived listener effort as an outcome measure," to be submitted for publication.

Chapter 4: Nagle, K.F., Isetti, D.I. & Eadie, T.L. "Everyday listener impressions of perceived effort when listening to electrolaryngeal speech," to be submitted for publication.

A summary table of this research follows on the next page.

Summary Table

Title	Purpose	Speech samples	Listeners	Listener task	Results
Ch.1: Objective versus subjective methods of measuring listening effort: A literature review	Review literature examining listening effort, with emphasis on subjective measurement of perceived listener effort (PLE)and disordered speech	NA	NA	NA	Effects on listening effort of disordered speech measured either objectively or subjectively are not well understood. Perceived listener effort measures may be uniquely sensitive to disordered speech.
Ch.2: Perceived listener effort for highly intelligible tracheoesophageal (TE) speech	Initial investigation of uniqueness of PLE compared to acceptability (ACC) with intelligibility controlled	2nd sentence of the Rainbow Passage produced by 14 male TE speakers	20 (8 male) inexperienced listeners, ages 18-32	Rate paired samples on 100 mm visual analog scale (VAS) for PLE and ACC across 2 sessions	Very strong negative correlation (r = -.99) between PLE and ACC. Individual rater variability in sample ratings across dimensions suggesting conceptual difference between ACC & PLE.
Ch.3: Utility of perceived listener effort as an outcome measure for disordered speech	Investigation of uniqueness of PLE compared to acceptability and intelligibility, and of effects of sentence length on all measures.	45 to 50-sentence sets of Short (5-7 words) and Long (13-15 words) sentences produced by 10 healthy males using an EL	25 inexperienced listeners; eventually reduced to 11 listeners (6 male) ages 18-35	Transcribe and rate a set of sentences for PLE and ACC using VAS, across 2 sessions	Very strong and unique relationship between PLE and ACC. Length x intelligibility interaction for PLE not present for ACC ratings, suggesting ACC is rated as a whole, whereas PLE ratings vary as a function of variables such as length.
Ch.4: Everyday listener impressions of perceived effort when listening to electrolaryngeal (EL) speech	Exploration of everyday listeners' perceptions of their effort when listening to disordered speech and their interpretations of the concepts of PLE and ACC	Same as Chapter 3	25 listeners; interviews from 14 listeners (7 male), ages 18-35, reviewed before conceptual saturation	Participate in cognitive interview following rating sessions	Listeners found EL speech difficult to understand. They sympathized with speakers, but did not like the sound of the speech. Experience with EL speech improved their ability to understand, but also increased fatigue. PLE and ACC are largely influenced by understandability. Some listeners distinguished PLE from intelligibility.

ACKNOWLEDGMENTS

First, I have been exceedingly fortunate to have collected such an esteemed supervisory committee. Tanya Eadie, you have been the best mentor I could have asked for. My words do not work to express my gratitude for all of your help and generosity. I have appreciated your willingness to let me take my time digging around in other yards, while keeping me on a long leash. I am honored to have been your first doctoral student, but more importantly, I am pleased to be your friend. Thank you to Debbie Kartin for your guidance during my fellowship and as a member of my committee. Kristie Spencer, thank you so much for listening, questioning, and always providing wise counsel. You have been a calming force throughout, and I would gladly sit on your floor any time. Huge thanks to my co-author and committee member Kathy Yorkston, for helping me find the big picture and being such an all-around help academically and professionally. To Richard Wright I owe a great debt for allowing me into that first seminar on vowel perception, including me in the Phonetics Lab and ultimately agreeing to be a part of this.

Lesley Olswang, you have been a model of good sense, good teaching and general awesomeness. I have been so lucky to have your mentorship for these many years. Jessica Sullivan patiently and generously let me use her lab. Thank you also to early committee members who have moved on: Brian Dudgeon and Chris Stecker always treated me as if they believed I knew what I was talking about, and it was appreciated! To our graduate program coordinator Ernie Lefler, I am pretty sure I could not have gotten through without you. Thank you to the many research participants and my many assistants in the Vocal Function lab: Heidi, Reyhaneh, Devon, Jody, Arlene and Erena. Your work has been invaluable – I hope you enjoyed it!

Michael Cannito graciously provided data and guidance on methodology I used for my dissertation experiments, and Megan McAuliffe allowed me to pick her brain not once, but three times, when I was planning them. My grand-mentor, Philip Doyle has been generous and supportive over the years and across the miles. Cara Stepp has been my go-to for a wide variety of questions arising in the later phases of writing this document. She has truly kept me going.

Everything was made easier and better by my dear friends from the department - you know who you are! And my friends from the outside world - Mark, Shaye, Heather, Tammy - listened to me and ate cupcakes with me and occasionally helped me waste my time very productively. I owe extra-large thanks to Violet and Mike from Chocolati Café for providing me with a sympathetic ear, an office away from school and a multitude of delicious mochas.

To my family go my greatest thanks. Knowing my cousins and my dear aunt were so close has been a comfort and a pleasure. My “brothers” Derek, Mike, Jay and David have meant more to me than they can know. Matilda, my spirit animal, has regulated my heartbeat with her steady purr. My amazing and accomplished sisters have provided a ridiculously high standard for me to meet. And to my wonderful parents, your generosity, thoughtfulness and support have sustained me. I promise this will be the last graduation!

DEDICATION

to anyone who has struggled to understand or to be understood,
for whatever reason

Introduction:
Statement of the problem

The chapters within this document trace a program of research on the effects of listening to disordered speech and voice for individuals lacking experience with such speech. The impetus for this research was the acknowledgment of the potential differences between experts' perceptions of disordered speech and everyday listeners' perceptions and assumptions. Specifically, erroneous assumptions about individuals with disordered speech may be based on difficulty understanding, rather than the perceptual qualities of the speech signal *per se*. If speech is too hard to understand, listeners may choose to avoid or cut short communicative interactions with the speaker; the listener may not want to carry the increased burden of processing and responding to disordered speech (Lindblom, 1990). This burden has been acknowledged and studied, but inconsistency in terminology and methodology have complicated the development of an integrated model (McGarrigle, Munro, Dawes, Stewart, Moore, Barry et al., 2014). This document has two main purposes:

1. Chapter 1 represents an attempt to summarize and compare the literature examining listening effort as an objectively measured concept and perceived listener effort (PLE) as experienced (and rated) by inexperienced listeners. The intention of this chapter is to differentiate *listening* effort from PLE, and to review the small number of studies that have investigated either concept for disordered speech.
2. Chapters 2 through 4 examine the relationships among intelligibility, speech acceptability and PLE, specifically for two types of alaryngeal speech.

First, Chapter 2 examines the very strong inverse relationship between PLE and acceptability ratings of tracheoesophageal speech using a paired comparison paradigm. By providing an external standard against which to rate each sample, the paired comparison paradigm maximizes listener reliability. Samples obtained with the same elicitation stimulus

were used to remove the process of decoding each sample from the task. The intentions of this study were to reveal whether there were differences between ratings of PLE and acceptability at roughly equal intelligibility, and to demonstrate that inexperienced listeners could reliably rate PLE.

The study reported in Chapter 3 took the further step of comparing PLE and acceptability to samples of disordered speech (in this case, electrolaryngeal) with a wide range of intelligibility. To show that PLE ratings might add unique outcome information despite high correlation with acceptability measures, the effect of sentence length was also tested for these measures. Chapter 4 contains a qualitative analysis of interview data obtained from the listeners who rated the samples in Chapter 3. Interviewing participants allowed them to comment on their observations about the speech samples and the rating process itself; in this way, data unlimited by researcher-imposed quantitative scales could be obtained. Everyday listeners reported how they interpreted the concepts of PLE and acceptability. They also described their impressions of PLE and acceptability for electrolaryngeal speech.

Through these chapters, PLE emerges as a unique outcome measure that may be sensitive to change and difference in disordered speech beyond other auditory-perceptual measures and objective measures of listening effort.

Chapter 1

Objective versus subjective methods of measuring listening effort: A literature review

Abstract

Research into the burden of listening to speech in adverse conditions is often conducted using a dual-task paradigm, in which listening effort (LE) is indexed by performance or reaction time on a task. This research is reviewed briefly, with an emphasis on the findings and methodology of recent work exploring variables related to the environment and listener characteristics. The evidence suggests that PLE and LE are different constructs; in other words, reductions in performance or processing speed are not the same as perceptions of working very hard. Studies examining the effects of a degraded signal (i.e., different and disordered speech) on perceived listener effort (PLE) are summarized in the interest of proposing the use of subjective ratings of listener effort for disordered speech. This literature review provides a rationale for further investigating the use of PLE as an outcome measure for disordered speech and voice.

1. Introduction

A communication disorder can have devastating effects on quality of life (QOL), affecting personal and work relationships and depriving affected individuals of their ability to express themselves verbally. Accurate diagnosis and evaluation of treatment outcomes are critical aspects of planning and documenting change, but can be complex for speech and voice disorders[1]. Multidimensional assessment of speech and voice may include acoustic, aerodynamic and other physiological measures, along with measures of speech intelligibility[2] and patient-reported outcomes. The gold standard of speech and voice outcome measures, however, is perceptual. In the absence of measurable anatomical or physiological signs, expert ratings of perceptual dimensions such as roughness, naturalness or "overall severity" can provide a sensitive marker of change or difference (Dagenais, Brown & Moore, 2006; Sussman & Tjaden, 2012).

It is not uncommon for two equivalently intelligible speech samples to differ drastically in how they sound. That is, speech that is 100% intelligible may have perceptual features that make it more distracting or difficult to process than other equally intelligible speech. Likewise, auditory-perceptual judgments of unintelligible speech may prove more functional than measuring the number of words correct. For these reasons, a combination of clinical outcome measures should predict communicative success; a speaker whose speech receives high ratings of acceptability or naturalness (or low ratings of roughness or breathiness) may be expected to perform relatively better in a communicative dyad than one with opposing ratings. All other things being equal, the more severe the communication disorder, the more demanding it may be

---

[1] In the interest of brevity, "speech disorder" will hereinafter refer to speech and/or voice disorder.

[2] NB: For the purposes of this paper, "intelligibility" will refer to speech intelligibility, or percent words correctly identified by a listener. In hearing research, intelligibility is a characteristic of the listener. "Listener intelligibility" will be used in this review when it can be determined that a given measure applies to a listener, not a speech sample.

for a listener to participate in a conversation with the individual with the speech disorder. However, the auditory-perceptual effects of speech impairment on unfamiliar communication partners are only peripherally considered when evaluating speech outcomes.

Communication partners are a major contextual factor in the communicative success of an individual with a speech or voice disorder (Eadie, 2007; O'Halloran, Hickson & Worrall, 2008). However, measures of communicative success such as the Communicative Effectiveness Survey (CES; Donovan, Kendall, Young & Rosenbek, 2008) take into account only the perspective of the individual with the communicative disorder. Of course, this individual is of primary concern. Whether the impairment is receptive or expressive, or both, it is the nature of communication that at least one other person be involved. There are implications for both parties in a communicative dyad, whether the healthy partner[3] is family, a caregiver or a complete stranger. Both people in a communicative dyad are equally responsible for the success of the interaction. When one partner fails to fulfill his/her role, the other may have to shoulder more responsibility for communicative success. When one of the partners has disordered speech or voice, this burden may be significant for both parties (Lindblom, 1990; Yorkston, Klasner & Swanson, 2001). The speaker may struggle to make his/her speech clear, while the listener strains to understand. If the burden on either becomes too high, the listener may withdraw from the interaction and avoid similar interactions in the future.

The International Classification of Functioning, Disability and Health (ICF) model (WHO, 2002) indicates that a major goal of intervention should be decreasing activity limitations

---

[3] Roles obviously switch throughout a communicative interaction. For simplicity, the individual with a speech or voice disorder will hereinafter be referred to as the speaker and the communication partner will be referred to as the listener.

and participation restrictions on the individual with the health condition, in order to maximize QOL. Activity and participation levels of functioning may not be observable in clinical situations and may be difficult to address directly. Similarly, behaviors observed in the speech-language pathology clinic may not carry over to the real world. Yet efforts to address the frequent disconnect between performance within and outside of a clinic have yet to include a measure directed at the increased burden of impaired speech or voice on listeners of any type. If individuals with a speech disorder desire to participate in communicative activities, it is critical to consider the auditory-perceptual effects of their speech on their "real-world" listeners.

1.1 Listener burden/listening effort

The concept of listener burden is not new to researchers in communication disorders. For example, hearing scientists are keen to know how to adjust speech processing algorithms used in hearing aids to reduce listening effort in hearing-impaired individuals. One issue with listener burden is that individuals experience it differently. This makes it difficult to measure reliably and complicated to apply research findings to any one listener. A related problem is the perplexing number of ways in which the concept has been defined, labeled and measured. As outlined in a recent "white paper" on the measurement of listening effort and fatigue, concepts related to listener burden have been investigated using multiple terms with multiple different meanings (McGarrigle, Munro, Dawes, Stewart, Moore, Barry et al., 2014). The authors tracked, reviewed and grouped research on listening effort and fatigue based on methodology, with the goal of identifying what was actually measured in each study. As identified by the authors, the term "listening effort" was used to index perceived listener effort, auditory processing speed, the dual-task cost of listening on cognitive processes and the "physiological correlates of change in listening demand" (p.3). This is a wide range of meaning for one term. Conversely, specific

physiological correlates of change in listening demand had been labeled in the literature using the following variety of terms: cognitive effort, effortful listening, processing load, listening effort, resourceful listening, cognitive load and listening difficulty. The context for discussing listener burden is thus complicated by a lack of agreement on terminology coupled with a shaky conceptual foundation. Even the attempt of McGarrigle et al. (2014) to provide a standard definition of listening effort as "the mental exertion required to attend to, and understand, an auditory message" (p. 2 early online) leaves open questions about the operational definitions of exertion, attention, understanding and auditory messages. For example, would attention without comprehension count as listening effort? Is mental exertion a purely objective term?

Adverse conditions related to noise, degraded speech and hearing loss are known to affect speech recognition. In a review of speech recognition research, Mattys, Davis, Bradlow and Scott (2012) classified adverse conditions as originating either internally or externally to the listener. They grouped adverse conditions further into three categories: source/signal degradation, environmental or transmission degradation and receiver limitations. The bulk of research on listening effort explores factors inherent to listeners (i.e., receiver limitations such as hearing impairment, processing speed, working memory capacity) and to variables within the listening environment (i.e., environmental/transmission degradation such as noise level, noise type, availability of visual cues). Hearing-impaired listeners tend to expend more objectively measured listening effort than normal-hearing listeners under similar conditions (Desjardins & Doherty, 2013; Larsby, Hallgren, Lyxell and Arlinger, 2005). Providing visual cues may reduce listener effort particularly for listeners with high lip-reading ability (Picou, Ricketts & Hornsby, 2011; 2013). Informational masking noise (i.e., carrying interfering linguistic information) has consistently been found to increase listening effort compared to energetic masking noise

(Brungart, Iyer, Thompson, Simpson, Gordon-Salant, Shurman et al., 2013; Koelewijn, Zekveld, Festen & Kramer, 2012). Higher signal-to-noise ratio (SNR) predictably reduces listening effort (Desjardins & Doherty, 2013; Zekveld & Kramer, 2014). Listening effort during hearing aid use can also be differentially affected by adjustments to dynamic compression, noise reduction and other speech processing algorithms (Brons, Houben & Dreschler, 2013; Gustafson, McCreery, Hoover, Kopun & Stelmachowicz, 2014).

Very few studies have examined the burden associated with a speech signal that is itself distorted. In other words, the effects of disordered speech or voice have been largely overlooked, despite their clear potential to increase listening difficulty. Information about the effects of disordered speech or voice is vital when predicting the success of a communicative interaction involving an individual with a speech, voice or hearing impairment. Thus, the goal of this paper is to provide context and a rationale for investigating perceived listener effort for disordered speech and voice.

First, existing listening effort (LE) research will be reviewed briefly, with an emphasis on the findings and methodology of recent work exploring variables related to the environment and listener characteristics. As described in McGarrigle et al. (2014), much of this research seeks to report an objective measure of LE by instrumentally quantifying reductions in the accuracy or speed of speech processing. Findings pertaining to LE are reported in terms of relative performance or speed on two related tasks (i.e. using a dual task paradigm), or as changes in a physiological measure such as stress hormone levels or pupil dilation.

Research on *perceived listener effort* (PLE), on the other hand, uses data obtained through subjective reports and ratings often generated by listeners unfamiliar with degraded speech. Because of the inter-listener variability inherent in perceptual scaling, the objectivity

captured in some psychophysical or physiological measures of listening effort cannot be achieved by obtaining ratings of PLE. Nonetheless, there is potentially a great deal to learn about the way both normal-hearing and hearing-impaired listeners perceive the amount of effort they expend on listening to speech that is different (i.e., accented) or disordered. As hearing-impaired listeners may perceive increased effort to process normal speech because of a fault in the "receiver," so might normal-hearing listeners be aware of increased effort to process disordered or different speech because of a degraded "signal." Because of the dual role of partners in a communicative dyad, the impression of the listening burden in both cases is relevant to the communicative success of individuals with either speech or hearing impairments.

Self- (i.e., listener-) reported measures of LE are generally obtained in hearing clinics, using some type of rating scale (Hornsby, 2012; Humes & Humes, 2004; Noble & Gatehouse, 2006). They are easier and cheaper to implement than more objective measures requiring instrumentation. More importantly, there have been continued conflicting findings about the relationship between objective and subjective measures of LE (Desjardins & Doherty, 2013; Gosselin & Gagné, 2011; Larsby et al. 2005; Picou et al., 2011; Zekveld & Kramer, 2014), suggesting that objective and subjective measurement are indexing conceptually different phenomena. Therefore, the second part of this paper will review the few examinations of the effects of a degraded signal (i.e., different and disordered speech) on PLE. The paper will culminate with an organizational framework for the proposed series in this dissertation that relates to PLE.

2. Environmental and Internal Factors

Searches in Web of Science and PubMed were conducted to locate relevant literature for this review using the following related terms: *listening effort, listener effort, attention allocation,*

perceived listening effort, perceived listener effort, cognitive load, processing load, listening ease, listening difficulty, listener burden, perceptual effort. Papers were then reviewed to determine their relevance to the topic of listening effort and its measurement, with an emphasis on the association between objective and subjective measures. A summary of relevant findings is reported in the following review.

According to McGarrigle et al. (2014), the literature investigating LE and closely related concepts such as perceptual effort and processing load can be divided into three methodological categories, which are reviewed in the following pages. First, behavioral methods measure LE in terms of changes in auditory processing speed and performance on increasingly challenging listening tasks. Second, physiological effects of challenging listening tasks have been measured in the nervous, muscular, circulatory and endocrine systems (e.g., Bernarding, Strauss, Hanneman, Seidler & Corona-Strauss, 2013; Hicks & Tharpe, 2002; Mackersie & Cones, 2011; Winn, Litovsky & Edwards, 2014). Third, listener-reported measures of PLE are typically obtained using perceptual scales and questionnaires. Behavioral and physiological outcomes have been interpreted as indexing LE in a relatively objective way (i.e., LE), whereas listener-reported measures of LE convey the perceived effort of a listener as he or she subjectively experiences it (i.e., PLE). Behavioral and physiological outcomes have often been obtained in research without input or comparison to perceived effort (e.g., Bernarding et al, 2013; Choi, Lotto, Lewis, Hoover & Stelmachowicz, 2008; Houben, van Doorn-Bierman & Dreschler, 2013). This may be important, as many of the studies that have tested both objective and subjective measures of listening effort have found no correlation between them (Downs & Crum, 1978; Feuerstein, 1992; Fraser, Gagne, Alepins & Dubois, 2010).

2.1 The dual-task paradigm

Because they frequently involve listening tasks, behavioral studies of cognitive load, working memory and attention have provided some information about LE. These studies often use a dual task paradigm in which performance on a secondary task is influenced by making a primary task more challenging. Briefly, the amount of effort required to complete a primary task (usually speech recognition), is measured by adding a secondary interfering task (see review articles by Gosselin & Gagne, 2010 and McGarrigle et al., 2014). Adding complexity (e.g., noise, increasing stimulus length) to the primary task leads to decreased performance on a simultaneous secondary task. This decreased performance signifies increased effort. In this type of study, adding complexity to the primary task is meant to stretch the limits of the cognitive system so that it is at maximum capacity, but still capable of performing the primary task. As the load induced by the primary task increases from "simple" to "difficult," reaction times for the secondary task should increase or performance should decrease, providing an objective indicator of cognitive effort. For example, when the secondary task is to respond immediately to a light probe, probe reaction time should increase when noise is added to the primary task, indicating that the cognitive system is being stressed (Downs & Crum, 1978; Downs, 1982; Feuerstein, 1992; Hicks & Tharpe, 2002). A basic assumption in this type of research is that the cognitive system has limited resources available at any one time, and that devoting proportionally more resources to one task will reduce the resources available for the second (Kahneman, 1973). Further, complex tasks involving working memory may reveal the cognitive processes activated when processing a degraded signal (Rönnberg, Rudner, Foo & Lunner, 2008; Rönnberg, Lunner, Zekveld, Sörqvist, Lyxell, Dahlström et al., 2013).

Designing a dual task experiment requires careful consideration of the type of processing resources used in accomplishing both tasks (Fisk, Derrick & Schneider, 1986). As described above, single and dual *primary* task performance must be equivalent, or secondary task performance in the "easy" versus "difficult" dual tasks would be uninterpretable. Subjects must be instructed to focus on the primary task at all times. Hypothetically, when the secondary task is added to the first, remaining cognitive capacity should be devoted to the primary task; secondary task performance is meant to suffer in the dual-task experiment, relative to single task performance. Second, all tasks must draw from the same pool of resources so that the capacity of that system is taxed enough to affect secondary task performance when the primary task becomes more difficult. In other words, the secondary task must actually interrupt primary task processing. Finally, the secondary task must require a constant cognitive load of its own throughout the experiment, so that learning-based improvement does not interfere with interpretation of reaction time differences. Whatever the secondary task is, it should not improve with practice; it should require controlled or effortful processing subject to capacity limits (Fisk et al., 1986).

Results of research using dual tasks show an important but unsurprising role of working memory in performing challenging tasks. For example, Amichetti, Stanley, White and Wingfield (2013) used a word recall task to assess listeners' ability to judge the limits of their own working memory. They tested the effects of adding a challenging task to the baseline recall task, by making listeners interrupt experimental trials based on estimating their own ability to recall; listeners were to suspend presentation of a word list at the latest point at which they would be able to recall the words with 100% accuracy. As conditions grew more effortful (i.e., when listeners had to choose when to stop), listeners who heard the stimuli at lower SNR were

significantly less able to accurately monitor their working memory capacity than those who heard it at a higher SNR. Studies like this support the view that cognitive resources are limited and must be shared among tasks (Baddeley, 2000; Kahneman, 1973).

2.2 Dual-task measurement of listening effort

The previous studies provide insight about models of working memory capacity and attentional focus, not on listening itself. Another set of literature investigates listener burden specifically using dual task experiments. Gosselin and Gagné (2010) reviewed nine studies of LE conducted between 1958 and 2008, in which the primary tasks were recognizing words or speech in noise. Some secondary tasks were measured in terms of accuracy of single word recall, visual tracking or tactile pattern recognition; however, the most common secondary tasks measured LE using processing speed, or response time to a probe. Examples include measurements of reaction time to a light probe (Downs & Crum, 1978; Downs, 1982; Hicks & Tharpe, 2002) and serial digit recall (Rakerd, Seitz & Whearty, 1996; Choi et al., 2008).

Measuring processing speed as an index of LE may be most useful in conditions of highly intelligible speech. An outcome measure that is sensitive in conditions of perfectly intelligible speech would be valuable for describing speech that is clearly different from typical speech. If all the words in an utterance have been accurately identified, presenting the utterance at a louder level or reducing the amount of background noise could not make it more than 100% intelligible. If processing speed increases in these conditions, the difference can be interpreted as an increase in processing efficiency. This increased efficiency would in turn be interpreted as an objective indicator of decreased LE. Houben, et al. (2013) measured differences in response times to two tasks in optimal listening conditions (i.e., at almost 100% intelligibility). Using normal-hearing listeners, they found that processing speed continued to increase as SNR

increased, suggesting that listeners took advantage of increased access to acoustic cues to listen more efficiently. These findings support the view that measures of LE may provide information unavailable from other outcome measures. Specifically, for equally intelligible samples, objective measures of LE may prove sensitive enough to reveal differences that would otherwise not be found.

In addition to working memory capacity, internal listener characteristics such as receptive vocabulary, hearing status, age and motivation have also been found to affect speech recognition and LE in adverse listening conditions. For example, Tamati, Gilbert and Pisoni (2013) examined differences among listeners based on their performance on a speech recognition task presented in multitalker babble noise (Perceptually Robust English Sentence Test Open-set; PRESTO). The authors measured "real world hearing" (questionnaire), indexical processing abilities (gender and talker discrimination and regional dialect categorization tasks) and neurocognitive processing skills, including working memory capacity (auditory digit span), attention/inhibition, vocabulary size, self-rated executive function and nonverbal IQ. The HiPRESTO group had better working memory capacity and receptive vocabulary and more accurately identified gender and dialect of the speakers than the LoPRESTO group. The groups were not otherwise statistically different. These findings point to a clear role for working memory (and receptive vocabulary) in speech recognition in adverse conditions. They also show that normal-hearing listeners can vary in a number of meaningful ways that could affect how group data on auditory-perceptual tasks are interpreted.

The effects of age and hearing status on LE have been demonstrated repeatedly in tasks measuring processing speed and accuracy (e.g., George, Zekveld, Kramer, Goverts, Festen & Houtgast, 2007; Hallgren, Larsby, Lyxell & Arlinger, 2005), but the relationship is not always

straightforward. For example, Larsby, Hallgren, Lyxell and Arlinger (2005) presented three tests of verbal information processing to young normal-hearing (YNH), older normal-hearing (ONH), younger hearing-impaired (YHI) and older hearing-impaired (OHI) listeners. All tests were presented in text-only, auditory-only and audiovisual modalities, each in three types of masking noise and quiet. Older and hearing-impaired listeners in general were less accurate and had longer reaction times than younger and normal-hearing listeners on verbal processing tasks. All listeners depended more on visual cues in noise than in quiet.

Gosselin and Gagné (2011) used a closed-set auditory word recognition test as a primary task and a vibrotactile recognition test as a secondary task in their investigation of the effects of age on LE. Older adults were significantly less accurate and more slow to perform the secondary task than the younger adults despite similar performance on the tasks in isolation, indicating that the older adults expended more LE. At equivalent performance levels (i.e., lower noise for older listeners), older listeners' ratings of PLE were significantly lower on a visual analog scale (VAS; 0 = negligible amount of effort) than those of younger listeners. At equivalent SNRs, mean PLE ratings provided by older listeners were the same as those from younger listeners even though the older listeners had longer reaction times. In other words, although the older listeners appeared to expend more effort than younger listeners based on their reaction times, their perceived effort was lower. In fact, the authors reported that ratings of PLE did not correlate with any of the dual task measures. They concluded that objective dual task measures and subjective ratings of PLE address different aspects of listening effort regardless of the age of the listener.

These investigations of the effects of age on LE have revealed predicted decreased performance and increased processing speeds for older listeners; however, they have also apparently uncovered an effect of increased tolerance or practice with degraded speech that

seems to lead older listeners to unpredictably lower ratings of PLE than their objective performance would suggest. This increased tolerance is consistent with reported higher mean ratings of acceptability of alaryngeal speech from older listeners compared to younger listeners (Law, Ma & Yiu, 2009). All of these are factors that must be considered when designing studies about PLE in disordered voice and speech.

An interesting contrast to these findings can be found in the results of a study of older hearing-impaired listeners, in which low-context stimuli were provided by younger versus older speakers (McAuliffe, Wilding, Rickard & O'Beirne, 2012). This naturally degraded speech included age-related differences in speech rate, pitch and pitch range. Listeners repeated the phrase stimuli and then marked the amount of effort required to recognize each phrase using a 10cm VAS. Although performance on the repetition task did not differ based on speaker group, listeners rated samples produced by older speakers as requiring significantly more effort overall than samples produced by younger speakers. Given numerous previously reported findings that older listeners rate their perceived effort rather generously, the authors suggested that the acoustic and perceptual differences between samples of younger versus older speech in this study must have been significant. Notably, however, the stimuli for this study were presented in quiet. Because increasing noise is generally used to increase the difficulty in listening tasks, it may be that the speech recognition task in this study was just not demanding enough to register performance differences in the presence of perceptual differences. These findings may indicate that ratings of PLE are more sensitive to differences between young and older speech than accuracy on a repetition task. How this might extend to ratings of disordered vs. typical speech needs further examination.

The role of motivation in attending to and expending effort on listening tasks has rarely been acknowledged by researchers; only recently has motivation been investigated as a variable of LE. Picou and Ricketts (2014) used passage pairs presented at two different SNRs to elicit ratings of PLE and "listening tiredness" on an 11-point equal appearing interval (EAI) scale. In the Low Motivation condition, subjects were told only to listen carefully to the passages; in the High Motivation condition, they were told to listen carefully and that they would be tested on the passages, for which they received immediate feedback. When presented without visual cues, motivation increased PLE ratings for only the lower SNR condition; listeners worked harder when necessary, possibly because they were motivated by knowing they would be tested. However, when visual cues were provided for the stimuli, listeners indicated increased effort at both levels of SNR regardless of motivation. The authors suggested that individual personality differences might have affected the results, but that the absent effect of motivation in the high SNR, auditory-only condition was probably due to the ease of this task (i.e., it did not require external motivation).

In summary, studies using psychophysical methods have shown that LE as measured by reaction times and performance measures is affected by working memory capacity, age, hearing impairment, noise type, noise level, modality (auditory-only vs. auditory-visual) and listener motivation.

2.3 Physiological methods

Listening effort has also been measured in terms of physiological changes that are known or believed to occur in conditions of increased mental effort, stress or fatigue. These types of studies use involuntary changes within listeners to measure the effects of externally controlled variables, such as task difficulty, noise level or noise type, on LE. Changes in central nervous

system activity during effortful listening have been measured using functional magnetic resonance imaging (fMRI), electroencephalography (EEG) and event-related potentials (ERPs). During challenging listening tasks, increased activity has been shown in specific areas of the brain aligned with memory and attention (Bernarding, Strauss, Latzel, Hannemann, Chalupper & Corona-Strauss, 2011; Bernarding, Strauss, Hanneman & Corona-Strauss, 2012; Bernarding et al., 2013). Findings thus far have provided further evidence of differences in LE for challenging listening tasks related to age and hearing impairment (for review, see McGarrigle et al., 2014).

Several labs have investigated LE via the autonomic nervous system, using pupil dilation as an index of focused attention or concentration (Kramer, Kapteyn, Festen & Kuik, 1997; Papesh, Goldinger & Hout, 2012; Winn & Edwards, 2013; Zekveld, Kramer & Festen, 2011; Zekveld, Festen & Kramer, 2013). Zekveld, Kramer and Festen (2010) specifically measured the effect of speech intelligibility level on pupil dilation, using adaptive procedures to find speech reception thresholds at 50%, 71% and 84% accuracy. They measured peak dilation amplitude, latency of peak dilation amplitude and mean pupil dilation during processing of speech in speech-shaped noise. They also obtained ratings of PLE using 9-point EAI scales marked at the odd numbers. As expected, mean ratings of PLE and all pupillometry measures decreased significantly as speech reception threshold decreased; however, inter-listener measures of these variables were not related. This study highlights two important factors of pupillometric research. Generally speaking, increases in peak dilation have been associated with more difficult tasks (i.e., listening with informational masking versus fluctuating noise with lower informational content; Koelewijn et al., 2012). Between listener differences in pupil responses, however, do not reflect differences in PLE or speech reception threshold. Changes in pupil dilation can be

complicated to interpret, but when pupil size is normalized, relative changes in pupil size seem to indicate the degree of mental effort required to process speech in noise.

Zekveld and Kramer (2014) measured cognitive load using pupillometry for sentences presented at three noise levels in each of four intelligibility conditions. Listeners also completed three 11-point EAI scales (e.g., PLE, estimated accuracy, and how often listeners gave up trying) after each of the 12 conditions. There was a significant difference in ratings of effort and performance for all conditions. Listeners reported "giving up trying" to perceive the stimuli more often in the medium- and low intelligibility conditions than in the high intelligibility and speech-in-silence conditions. Low peak pupil dilation was related to high ratings of giving up in the low condition. This is consistent with evidence that increased effort is demonstrated by increased relative pupil dilation. No other significant associations between PLE ratings and pupillometry were found in this study.

Physiological changes in the circulatory and endocrine systems as a result of increased LE have thus far failed to show any applicability for measuring LE. Mackersie and Cones (2011) measured skin conductance, skin temperature, heart rate and EMG activity on the frontalis muscle during increasingly challenging listening tasks. In this case, informational masking provided by competing talkers was meant to increase the difficulty of focusing on the target speech stream. The authors reported no significant change in performance on their speech recognition task across conditions, although skin conductance and EMG activity increased as task demand increased. Subjective ratings of effort were obtained using the NASA Task Load Index (Hart & Staveland, 1988), a 12 cm VAS. Ratings of PLE were strongly associated only with skin conductance ($r = 0.67$), and only at relatively high effort levels.

Hicks and Tharpe (2002) measured salivary cortisol levels in their investigation of the effects of hearing loss and fatigue on LE in school children (between ages 6 and 11). Cortisol is a hormone that increases in response to stressors; low cortisol levels are associated with fatigue. The authors hypothesized, but did not find, that children with hearing-impairment would have lower cortisol levels at the end of the school day, compared to normal-hearing children. Numerous reasons were provided for these results, including small sample size and the fact that the hearing-impaired children wore hearing aids during the school day. The authors also observed that what has been reported as fatigue may actually have been boredom or inattention. They suggested that salivary cortisol levels may not be sensitive enough to measure listening-related changes in stress levels.

Given the differences found in research between physiological and behavioral effects of effortful listening, it is extremely important to clarify what is meant by "listening effort." For research purposes, it may be possible to define LE as an increase in electrical activity in specific areas of the brain during a specific task, but clinical use of this method is not feasible. In fact, it is not yet known whether behavioral and physiological findings agree with what is arguably more important in the real world – the perceptions of the listener. Psychophysical and physiological methods all currently have similar drawbacks with respect to clinical feasibility and ecological validity, although improvements in technology may produce viable clinical options in the future (e.g., pupillometry; Winn, et al. 2014). If, as Bernarding et al. (2013) have reported, processing speed is *not* a good index of LE, questions remain as to what will ultimately provide the best and most accurate objective measure of LE. As noted by McGarrigle et al. (2014), the field will first have to come to some level of agreement on the operational definition of "listening effort."

2.4 Perceived listener effort

Research reviewed to this point has attempted to find objective indications that reductions in processing speed or task performance vary in predictable ways based on the difficulty of listening conditions. There is a briefer history of attempts to demonstrate associations between these LE measures and the effort listeners believe they have expended in performing those tasks. The *variables* affecting LE (i.e., working memory capacity, age, hearing impairment, noise) generally have similar effects on PLE, despite the weak relationship generally reported between LE and PLE. Work examining relationships between performance and perceived effort is summarized in Table 1.1 and reviewed below. This is followed by a discussion reviewing the effects of variables on measures of PLE.

Table 1.1 Review of research comparing objective and subjective measures indicated by "scale type" of listening effort.

Key to abbreviations
LE Measure: RT = reaction time
Scale Type: DME = direct magnitude estimation, EAI = equal appearing interval, VAS = visual analog scale
Stimulus type: A = auditory only, AV = auditory + visual
Listeners: OHI = older hearing-impaired, ONH = older normal hearing, YHI = younger hearing impaired, YNH = younger normal hearing

Title/Authors	Year	Terminology used for PLE	Objective LE Measure	Scale type	Stimulus type	Listeners	Correlation with LE
Downs & Crum "Processing demands during auditory learning under degraded listening conditions"	1978	learning ease	word recognition accuracy	7 pt EAI	spondee pairs in 8-talker noise	YNH	**Not significantly correlated**
Munro & Derwing, "Processing time, accent and comprehensibility in the perception of native and foreign-accented speech"	1995	comprehensibility (a.k.a., "ease of understanding")	intelligibility (accuracy) & true/false decision response latency	9 pt EAI	6-word questions produced by non-native speakers	YNH	Did not calculate association with response latency, but noted 5 point range of comprehensibility for equally intelligible (100%) samples
Hicks & Tharpe, "Listening effort and fatigue in school-age children with and without hearing loss"	2002	listening ease, fatigue	reaction time to probe light & cortisol level for fatigue	5 pt EAI	single words in 3 levels of 20-talker babble and quiet	NH & HI children	Reaction time difference between HI and NH **not seen in PLE** **No difference in cortisol level change between NH and HI**

Title/Authors	Year	Terminology used for PLE	Objective LE Measure	Scale type	Stimulus type	Listeners	Correlation with LE
Larsby, Hallgren, Lyxell & Arlinger, "Cognitive performance and perceived effort in speech processing tasks: Effects of different noise backgrounds in normal-hearing and hearing-impaired subjects"	2005	perceived effort	verbal information processing (accuracy) & RT	11 pt EAI	visual (text), audio and audio-visual in noise: 1. speech-shaped, 2. Hagerman's, 3. competing talker, and quiet	YNH, ONH, YHI & OHI	HI performed worse than NH and had higher PLE ratings. **ONH, OHI performed worse than YNH, YHI but did not have higher PLE ratings**
Panico & Healey, "Influence of text type, topic familiarity, and stuttering frequency on listener recall, comprehension, and mental effort"	2009	perceived mental effort	recall & comprehension accuracy	9 pt EAI	dysfluent read speech at 4 levels of severity; expository vs. narrative	YNH	Lower effort for greater fluency in general
Cote-Reschny & Hodge, "Listener effort and response time when transcribing words spoken by children with dysarthria"	2010	listener effort	intelligibility (accuracy) & response latency	DME	single words produced by children with dysarthria	YNH	Moderate correlation with response latency; weaker correlation with accuracy
Zekveld, Kramer & Festen, "Pupil response as an indication of effortful listening: the influence of sentence intelligibility"	2010	listening effort	pupillometry	9 pt EAI	sentences in noise	NH	**Not significantly correlated**

Title/Authors	Year	Terminology used for PLE	Objective LE Measure	Scale type	Stimulus type	Listeners	Correlation with LE
Fraser, Gagne, Alepins & Dubois, "Evaluating the effort expended to understand speech in noise using a dual-task paradigm: The effects of providing visual speech cues"	2010	listening effort	speech recognition accuracy & RT tactile pattern recognition accuracy & RT	100mm VAS	3-word sentences in 2 levels of broadband noise, A and AV	YNH	At equal accuracy, PLE ratings negatively correlated tactile task accuracy **in AV mode only**. At equal performance, significant correlation between PLE and both tactile measures **in A mode only**.
Picou, Ricketts & Hornsby, "Visual cues and listening effort: Individual variability"	2011	listening effort	word recognition accuracy	11 pt EAI	word recognition in quiet & 4-talker babble, A & AV	YNH	**No correlation reported**
Brons, Houben & Dreschler, "Perceptual effects of noise reduction with respect to personal preference, speech intelligibility, and listening effort"	2013	listening effort	intelligibility (accuracy)	9 pt EAI	low-context sentences in multitalker babble at 4 SNRs	YNH	**No correlation reported, but noted ceiling effect in PLE ratings**
Desjardins & Doherty, "Age-related changes in listening effort for various types of masker noises"	2013	"perceived ease of listening"	intelligibility (accuracy)	"restricted 0-100 DME"	in 2-talker, 6-talker and speech-shaped noise	ONH, OHI, YNH	**ONH reported more effort than OHI, despite similar performance** Significant correlation btw PLE and RT

Title/Authors	Year	Terminology used for PLE	Objective LE Measure	Scale type	Stimulus type	Listeners	Correlation with LE
Zekveld & Kramer, "Cognitive processing load across a wide range of listening conditions: Insights from pupillometry"	2014	listening effort	intelligibility (accuracy) & pupil dilation change	11 pt EAI	5-word sentences in competing talker noise at 4 levels of intelligibility	YNH	"Low intelligibility condition" (mean sentence intelligibility = 50% low SNRs) had higher PLE, but **no association with pupil dilation**

As has been shown for LE, studies have been performed showing how multiple variables affect PLE. For example, in a study investigating the effect of presentation mode and noise type on PLE, Larsby et al. (2005) also asked listeners to rate their "perceived effort" at the end of each condition. The authors described using a "combination of ratio and categorical scaling… ranging from 0 (none at all) to 10 (extremely great). If the experience was larger, the subject was allowed to use greater numbers" (p.133). Results pertaining to PLE are relevant here: there were main effects of hearing status, modality (auditory-only, auditory-visual and text only) and noise condition, but not of age. Although older listeners were less accurate and had longer reaction times, they did not report higher perceived effort. Further, there was a significant interaction between hearing status and modality; specifically, the hearing-impaired listeners perceived more expended effort for auditory-only than for audiovisual tasks. In addition to providing further evidence of the age effect on processing speed and accuracy, these results suggest that inclusion of visual information and reduction of noise are likely to reduce PLE.

Desjardins and Doherty (2013) used speech recognition in noise as the primary task and a secondary visual tracking task to investigate the effects of three background noise masker types on "perceived ease of listening" for younger, older and older hearing-impaired listeners. When performing the primary task, listeners also rated the ease of listening to the sentences in noise using a "restricted magnitude estimation scale," with 0 representing "very, very difficult" and 100 representing "very, very easy" listening. Speech recognition and listener ease were highest (i.e., most difficult) for sentences in speech-shaped noise. Although there were no differences in ratings of listening ease between the younger and older groups, the older normal hearing listeners provided lower ratings of listening ease than the older hearing-impaired listeners. Lower ratings on the scale used in this study meant that as a group, older normal hearing listeners found the

listening task more difficult than older hearing-impaired listeners. Listening effort was measured as time-on-target for the primary task in the dual task versus the primary task alone, and was similar for the older groups. These results demonstrate again that objectively equivalent LE may be accompanied by differences in PLE. Given that the hearing-impaired listeners rated samples as *easier* to listen to than the normal hearing older listeners, the authors suggested that experience with difficult listening conditions may have made the hearing-impaired listeners more tolerant of them. Thus, experience of the communication partner should be a consideration when evaluating effects of disordered speech and voice on PLE.

Hicks and Tharpe (2002) found a difference in LE between hearing-impaired and normal hearing children in a dual task paradigm with four noise conditions. As expected, response times were longer for hearing-impaired children, even though they wore hearing aids, than for normal-hearing children in all four noise conditions. Nevertheless, ratings of listening ease did not differ significantly between groups. The similarity of the hearing-impaired and normal-hearing groups' subjective ratings of listening ease in the presence of significantly different objective results is consistent with the potential "practice" effect posited for older listeners (Pichora-Fuller, Schneider & Daneman, 1995; Desjardins & Doherty, 2013).

As the findings of Larsby et al. (2005) and Picou and Ricketts (2014) might suggest, noise levels may mediate the benefits of access to visual cues during listening tasks. Two studies that measured PLE also examined auditory-only (AO) and auditory-visual (AV) modalities of stimulus presentation in noise. Fraser et al. (2010) reported that when the level of noise does not exceed the processing limits of listeners, they could take advantage of visual cues to improve performance and reduce PLE. When performance levels for the AO and AV modes were equalized (leading to an 8dB decrease in SNR for only the AV condition), listeners had to rely on

visual processing and process more noise than in the AO condition. Accuracy and processing speed were thus reduced because of the heavy processing load, and ratings of PLE in the AV condition were significantly higher than in the AO condition at equivalent performance levels. These findings were supported by Picou, Ricketts and Hornsby (2011), who investigated the impact of visual cues and noise on LE with a specific focus on individual variability. Immediately after a word recall task, listeners provided a subjective measure of PLE. They were asked to rate the degree to which they had to "put in effort to hear what was said," using an 11-point EAI scale. There was no effect of visual cues on word recognition, but there was a significant effect of noise on both word recognition and PLE ratings. Effects of visual cues on PLE were not reported.

Results of these studies are important for two reasons. First, the studies by Fraser et al. (2010) and Picou et al. (2011) also represent the few studies for which measures of both objective and subjective listening effort have been closely aligned. Second, they suggest that even useful cues such as visual information may increase processing load to a point beyond which performance is decreased, if the cognitive load is high for other reasons. The benefit of visual cues should not be taken for granted when stimuli are presented in noise; the same may be true when the stimuli themselves are "noisy," (i.e., degraded or disordered).

3. Source (signal) factors

To this point, research reviewed in this paper has tested normal or synthetic speech signals presented in varying types of noise at varying levels, occasionally augmented by visual information. However, a naturally degraded signal (e.g., apraxic, dysarthric or dysphonic speech) imposes a burden on listeners that has begun to be addressed using both quantitative and qualitative methods.

Klasner, Yorkston and Lillvik (2002) acknowledged the critical role of the listener in understanding dysarthric speech by analyzing data obtained from focus groups. In their study, everyday listeners heard speech samples produced by speakers with either Huntington's disease (HD) or amyotrophic lateral sclerosis (ALS). These listeners transcribed the samples, compared their transcriptions to target sentences and discussed what influenced their understanding of the samples. Mean sample intelligibility was 74% for the ALS speech (range = 62-89%) and 76% for the HD speech (range = 63-89%); samples were chosen to be difficult to understand. After qualitative analysis of the focus group data, the authors identified two themes (barriers and strategies) with four categories that described individuals' abilities to understand the sentences. Bottom-up barriers and strategies were related to segmental and suprasegmental aspects of dysarthric speech; top-down barriers and strategies were related to linguistic and cognitive aspects of understanding dysarthric speech. The authors hypothesized that the relatively consistent segmental and suprasegmental barriers to understanding observed in ALS speech led listeners to use bottom-up strategies, such as "I used the letters in the word that were clear to help me understand each word" (p. 296). Conversely, the inconsistent phonetic signal produced in HD speech led listeners to rely on linguistic and cognitive strategies for understanding, such as "I used the context of the sentence to understand the words that were not clear" and "I had to make a conscious effort to listen to every word to be able to understand the sentence" (p.296). Although all categories of strategies were used for both dysarthria types, differences in strategy for understanding ALS versus HD speech were also found in a follow-up study, in which a new group of listeners endorsed statements made by the focus groups (Klasner & Yorkston, 2005). Thus, everyday listeners described or endorsed barriers to understanding dysarthric speech, and explain or endorsed strategies for understanding it. The authors suggested that "better listeners"

may use different or more strategies to understand distorted speech. The characteristics of these "better listeners" are the subject of ongoing research (e.g., Tamati et al., 2013).

A few other studies have examined what happens to LE measures when the signal itself is degraded due to a communication difference or disorder, and their methodology differs from work reviewed in previous sections of this paper. These differences are important to consider, given that most of the prior research was focused on listeners with hearing impairment with little focus on the speaker or the nature of the speech samples. For example, Coté-Reschny and Hodge (2010) measured response time as time from auditory onset of the stimulus word to entry of the first letter of transcription of single-syllable words obtained from children with dysarthria. They reported a moderate correlation (rho = -0.57) between direct magnitude estimations of PLE and transcription accuracy. They also reported that the combination of response time and accuracy accounted for 60% of the variance in PLE scores.

Interpretation of these findings is complicated by the methodology used in this study. First, in choosing a modulus for ratings of listener effort, the authors chose a speech sample from a child whose overall intelligibility was 54%. This choice was undoubtedly motivated by the convention of using a modulus from the midpoint of the continuum being rated; however, use of a modulus from the continuum of intelligibility to obtain ratings of PLE implies that intelligibility and listener effort are actually the same parameter. The authors provide no description of the quality of the modulus itself; they report only the speaker's overall intelligibility. Second, the use of single-word stimuli limits the generalizability of findings. Single word intelligibility measures may provide important information for children with severe dysarthria, but any difference between intelligibility and PLE may be greater and more meaningful for longer, more complex stimuli. Third, the "accuracy" of stimuli was reported in

terms of the number of listeners who transcribed the words correctly. This is an unusual way to report intelligibility, which tends to be reported in percent words or phonemes correct, and likely the result of using single word stimuli. Finally, and most importantly, the listeners in this study heard the stimuli twice, first when they transcribed the words and again when they provided ratings of PLE. As the authors themselves pointed out, the separation of these tasks may have affected the relationship between them. Having heard a stimulus twice before rating PLE could have created a bias related to perceptual learning or exposure to the stimuli. Conversely, the temporal separation of the tasks may have weakened any correlation between accuracy and ratings of PLE that may have been revealed if they had both been obtained on the same presentation of a stimulus.

Munro and Derwing (1995) also investigated the phenomenon of PLE in the field of second-language acquisition (SLA), using the term "comprehensibility" of speech.[4] Comprehensibility in this study was defined as "listeners' perceptions of difficulty in understanding particular utterances" (p. 291). Listeners evaluated the truth value of statements produced by native speakers of either English or Mandarin, and then transcribed them. They also rated the comprehensibility and "accentedness" of the sentences on a 9-point EAI scale. Response latency was measured as the time taken to make the decision as to whether a given statement was true. As expected, low-comprehensibility ratings (i.e., 7-9) were related to significantly longer response latencies than moderate- or high-comprehensibility ratings (i.e., 1-

[4] This terminology may lead to confusion because in the speech disorders literature, "comprehensibility" refers to the overall ability of an individual to be understood using whatever tools and strategies are available to him or her (Yorkston, Strand & Kennedy, 1996). "Comprehensibility" is a multidimensional concept in both SLA and speech-language pathology; however, in speech-language pathology, it is meant to be used as an overall measure of communicative success. It is clear that listener-, speaker- and contextual factors play a role in either interpretation of "comprehensibility." The term "perceived listener effort" aims to focus primarily on the listener in relation to a speaker or sample, while the concept of "comprehensibility" primarily attaches to the speaker. To describe an individual's comprehensibility appropriately, one would need to consider visual and gestural cues and the speaker's use of strategies that facilitate communication.

6). Response latency as to truth value was longer for stimuli produced by Mandarin speaker than by English speakers, and 17 of the 20 listeners took longer to verify the statements produced by Mandarin speakers than by English speakers. Intelligibility was quite high among the stimuli used in this study; ninety-four percent of the utterances were transcribed accurately. This indicates that the sample set was highly skewed, yet the distribution of comprehensibility ratings was fairly evenly distributed across the full scale.

This research provides an example of naturally degraded (foreign-accented) speech affecting both LE and PLE, and provides further evidence that the sensitivity lacking in measures of intelligibility at very high levels of intelligibility may be differentiated by ratings of PLE. Caution is advised in interpreting these results, however, because ratings were obtained during the second or third presentations of the stimuli. Although there was no significant difference in ratings provided in the second versus third presentations, the fact that listeners had already heard the statements and transcribed them may have affected the relationship between their ratings and response latencies.

Finally, Evitts and Searl (2006) examined differences in reaction time and performance in identifying single-word stimuli as the same as or different from an orthographic stimulus presented on a computer screen. A single proficient speaker produced stimuli for each condition. The intent of the study was to examine differences in LE among laryngeal, alaryngeal and synthetic speech. They reported that synthetic speech required the most LE, followed by electrolaryngeal, esophageal, tracheoesophageal and normal laryngeal speech; the difference between tracheoesophageal and laryngeal speech was not significant. Strictly speaking, this study did not use a dual task paradigm. Arguably, the primary task was recognizing the auditory stimulus and the secondary task was deciding if it matched the word on the screen. The laryngeal

mode served as a baseline condition, and increased reaction times are consistent with the view that alaryngeal speech is harder to process. However, because auditory and written stimuli were combined in a single decision task, it is difficult to compare these findings to other investigations of LE.

3.1 Relation between PLE & intelligibility

To date, ratings of PLE for disordered speech have not been compared to psychophysical or physiological performance; however, five studies have measured the perceived effort required to listen to disordered speech against performance on transcription or comprehension tasks (i.e., intelligibility).

Whitehill and Wong (2006) investigated the "listener aspects of intelligibility" for disordered speech. Inexperienced listeners identified the perceptual features of dysarthric speech contributing most to their judgments of listener effort rated on a 10 cm VAS. Although there was a strong negative correlation between intelligibility and PLE ratings (Spearman's r = -0.95), there were three speakers who were highly intelligible, yet received moderate PLE ratings (greater than 5 on the 10 cm VAS; group M = 4.54, SD = 2.77). These results provide further evidence that ratings of PLE may supply unique information about communicative success beyond intelligibility scores.

Panico and Healey (2009) introduced a linguistic component to their investigation of perceived mental effort when listening to increasingly dysfluent speech samples. They tested effects of both the task type and stimulus severity. They asked listeners to rate the mental effort required to recall and comprehend speech produced by a single speaker at four levels of dysfluency (0%, 5%, 10% and 15% of words). After hearing four familiar and unfamiliar narrative and expository speech samples at a single level of dysfluency, listeners completed

recall and comprehension tasks and rated their "perceived mental effort while listening to the speech and the content of the sample" on a 9-point EAI scale. Mean mental effort ratings increased as dysfluency increased, but the main effect of level of dysfluency was not statistically significant. Interactions between the 0% level and other levels of dysfluency across "text type" suggest that once listeners detected any dysfluencies in a sample, they perceived expending more effort to listen to it. There was a significant effect of topic familiarity, and mental effort ratings were consistently higher for expository than for narrative samples. The authors attributed the latter to the relatively predictable organizational structure of narratives compared to expositional samples. This would indicate that linguistic content, not just linguistic level, may affect measures of PLE. Relations between speech quality and intelligibility and PLE may vary not only with the linguistic complexity of the task, but also with its content. As a result, linguistic complexity of the stimulus is another factor that needs to be considered when designing studies examining PLE in disordered speech and voice.

Beukelman, Childes, Carrell, Funk, Ball and Pattee (2011) used "self-perceptions of attention allocation" to label their measurement of auditory-perceptual load experienced for sentences produced by speakers with amyotrophic lateral sclerosis (ALS). Five inexperienced listeners transcribed the sentences and rated the amount of attention required to listen to them on a 7-point EAI scale. Results were reported as mean speaker scores and ranged from less than 1 ("attention allocation required when listening to a professional announcer over the radio") to nearly 7 ("attention required when listening to a very important message under extremely difficult listening conditions"). Mean intelligibility scores ranged from 3.6%-100%, although 82% (27/33) of speakers were at least 75% intelligible on average. There was a strong negative correlation between perceived attention allocation and intelligibility (r = -0.89), but the highest

ratings of attention allocation occurred for speakers with mean intelligibility scores between 75% and 85%. Conversely, mean attention allocation scores were much lower for speakers with lower (<70%) and higher (>85%) mean intelligibility. The authors concluded that the apparent nonlinear relationship between attention and intelligibility may indicate that listeners try less hard when speech is either easy or very difficult to listen to. There may be a peak of "attention allocation" that occurs in a sweet spot where listeners believe they have a chance at understanding the entire utterance; when intelligibility is higher than this peak, listeners pay less attention because they can. When intelligibility is low enough, they do not pay attention because it is too frustrating or unproductive to do so. This interpretation is consistent with Zekveld and Kramer's (2014) finding that listeners reported "giving up" more frequently in both the medium- and low-intelligibility conditions of their study than in the high-intelligibility condition.

Beukelman, Gillespie, Fager and Ullman (2014) used similar methods to evaluate "perceived attention allocation" for speech produced by speakers with traumatic brain injury. Mean intelligibility scores and ratings of perceived attention allocation were obtained from 30 listeners, based on sentences produced by 27 speakers. Mean intelligibility scores ranged from 3% - 97% and were more equally distributed than in the 2011 study using speakers with ALS. Although the specific association between attention allocation and intelligibility was not reported, the data appeared to be nonlinearly related. Below 70% mean intelligibility, mean attention allocation scores ranged from 6-7 on the 7-point EAI scale; however, above 70% mean intelligibility, perceived attention allocation scores were much less predictable. These results support the results of the 2011 study, and provide further evidence that intelligible speech may still require moderate levels of attentional effort to understand.

Finally, PLE has shown some promise as a sensitive outcome measure for foreign-accented speech. In a single-subject treatment study, phonomotor treatment developed for rehabilitation of anomia was implemented to improve the production of American English sounds by a native speaker of Mandarin Chinese (Oelke, Sachet, Nagle, Bislick, Brookshire & Kendall, *submitted*). As a measure of generalization of trained phonemes, sentence and discourse level intelligibility was obtained using three unfamiliar listeners blinded to test period (pre- versus post-training), who also provided ratings of their PLE on a 9-point EAI scale. Intelligibility at these higher linguistic levels did not change significantly, possibly because it was high in pre-training probes; however, sentence-level productions required significantly less perceived effort to understand after training than before ($p = 0.002$). Intra-rater levels of agreement were not reported for sentence-level productions, but inter-rater agreement with mean ratings was relatively strong (average measures intra-class coefficient = 0.70). Although only three listeners provided the data in this Phase 1 study, these results provide further evidence that ratings of PLE may provide a measure of needed sensitivity in the absence of a difference in intelligibility among samples.

3.2 Relation between PLE & auditory-perceptual dimensions of speech

Two studies have compared PLE ratings to perceptual measures of speech stimuli, namely acceptability of tracheoesophageal speech (Nagle & Eadie, 2012b) and overall severity of the speech associated with adductor spasmodic dysphonia (ADSD; Nagle & Eadie, 2012b). Both studies used a paired comparison paradigm in which each sample was paired with every other sample, necessitating use of the same elicitation stimulus for all. That is, all samples were productions of the second sentence of the Rainbow Passage (Fairbanks, 1960). In this way, intelligibility was "controlled;" listeners did not have to decode the samples. Perceptual

dimensions were rated on 100 mm VAS, with lower ratings indicating more effort, more severity and less speech acceptability. For tracheoesophageal speech, mean ratings of PLE ranged from 25-79 mm and were very strongly negatively correlated with speech acceptability ratings (r = -0.99), which ranged from 21-81mm. However, two listeners who were equally reliable as the rest of the study group showed different patterns in their ratings of samples depending on the dimension being rated. This result suggests that, for these listeners at least, there may be a meaningful difference between PLE and speech acceptability, beyond their strong inverse relationship. For ADSD speech, mean PLE ratings ranged from 25-68 mm and were very highly positively correlated with overall severity ratings (r = 0.98), which ranged from 23-71 mm. The wide range of ratings used for PLE, speech acceptability and overall severity provides further evidence that these measures are sensitive when other measures are not. Apart from the individual listener differences found in the (2012a) study of tracheoesophageal speech, however, these findings seem to suggest that ratings of overall severity and speech acceptability may capture the same perceptual aspects of disordered voice as ratings of PLE.

4. Conclusion

This literature review has provided a rationale for further investigating the use of PLE as an outcome measure for disordered speech and voice. As discussed by McGarrigle et al. (2014), the need for consistency in terminology for studies of listening effort has been highlighted, and an attempt has been made to separate and clarify "objective" findings on listening effort (LE) from subjective findings on perceived listener effort (PLE). Factors affecting both LE and PLE have been reviewed, and the relationship between LE and PLE has been summarized (see Appendix for proposed model). Research measuring PLE for a degraded signal (i.e., different or disordered speech or voice) has been examined to describe what is known about the relationships

between PLE and intelligibility and the auditory-perceptual parameters of speech acceptability and overall severity. Briefly, the bulk of evidence suggests that although PLE appears to be influenced by similar factors (working memory, age, etc.), ratings rarely correlate with objective measures of LE and their association with intelligibility is unclear.

There is a great deal of evidence implying that PLE and LE are different constructs; although clearly related, reductions in performance or processing speed are not the same as perceptions of working very hard, and the research reported here supports that view (Gosselin & Gagne, 2011; Zekveld et al., 2011; 2014). Whether PLE provides unique information as an outcome measure, however, remains a question for future research, and is the focus of the proposed set of studies in this dissertation. Recent findings of strong relationships between PLE and perceptual measures of speech suggest that acceptability may serve as an inverse proxy for PLE. Equally, overall severity of speech may provide quantitative measurement of a signal that matches PLE ratings well enough for clinical use. In other words, measuring an internal, self-focused listener perception and a signal/speaker focused dimension of speech may be making a distinction without a difference.

4.1 Program of research

The chapters within this document trace a program of research on the effects of listening to disordered speech and voice for individuals lacking experience with such speech. The impetus for this research was the acknowledgment of the potential differences between experts' perceptions of disordered speech and everyday listeners' perceptions and assumptions.

4.1.1 Chapter 2: Perceived listener effort for highly intelligible tracheoesophageal speech

If measures of overall speech severity, speech acceptability or both are interchangeable with PLE, there is little reason to pursue the use of PLE as an outcome measure for disordered

speech. Acceptability is commonly used as an outcome measure for alaryngeal speech (Bennett & Weinberg, 1973) and overall severity is even more widely used to describe disordered speech and voice of all other types (Bunton, Kent, Duffy, Rosenbek & Kent, 2007; Carding, Wilson, MacKenzie & Deary, 2009; Cullinan, Prather & Williams, 1963). However, there are numerous reasons why outcomes obtained for one of these dimensions might not be transferable to another. First, the terms themselves may represent different concepts to different listeners, as they are meant to. Overall severity and acceptability are meant to reflect qualities of the speech signal; although characteristics of the listener-rater may be inextricable from his or her rating of a perceptual concept, the focus should be outward, on the stimulus. Listening effort, however, has a nearly opposite focus. The characteristics of the stimulus and the environmental context will affect PLE ratings, but the focus should be inward, on individual effort. There is evidence that unfamiliar listeners can differentiate their interpretations of seemingly closely-related perceptual concepts. For example, Brons et al. (2013) found that ratings of "overall preference, speech naturalness" and "noise annoyance" for different noise reduction algorithms were not significantly correlated with ratings of PLE.

The use of terms such as "speech acceptability" and "overall severity" of speech may also differ among disorder types. Thus, it is possible that the relations among these concepts and PLE fluctuate when applied to different types of disordered speech or voice. As an initial step to examine these relationships in those with a communication disorder, the first experimental study (Chapter 2) was designed to examine relationships between acceptability and PLE in those who have undergone total laryngectomies and who use tracheoesophageal speech. In this study, intelligibility was controlled by obtaining stimuli using the same elicitation stimulus; with

intelligibility relatively constant, the specific relationship between acceptability and PLE could be tested.

It is not currently known how task performance on a relatively objective measure such as intelligibility relates to PLE compared to other perceptual dimensions, such as speech acceptability. As described above, PLE has been compared to intelligibility and to perceptual dimensions for degraded/disordered speech, but not simultaneously. Likewise, very little is known about the effects of stimulus length or linguistic complexity on PLE. As a result, research described in Chapters 3 and 4 of this document investigates these questions using both qualitative and quantitative methods.

4.1.2 Chapter 3: Utility of PLE as an outcome measure for disordered speech

As mentioned above, PLE may be nonlinearly related to intelligibility and ratings of PLE may be more useful at "perfect" intelligibility. As a result, the next study in this dissertation was designed to examine these relationships across a range of intelligibility levels. Use of an electrolarynx allowed collection of samples with a "natural" range of intelligibility and speech acceptability; no additional noise or speech synthesis was necessary to avoid ceiling effects.

Results from this review revealed several factors that can affect PLE, above and beyond factors related to the speech sample. For example, PLE has been investigated using only a few types of structured stimuli (i.e., known lists of words and sentences). Ratings of spontaneous speech quality, for example, may diverge from ratings of PLE because of the cumulative effects of listening to unknown and unpredictably long samples of disordered speech. For example, context and content variation could lead to similarly dissociated ratings of the signal and PLE (Panico & Healey; 2009). PLE could also increase as length or duration of utterances increases for samples of equivalent perceptual quality. The study described in this chapter therefore

addresses the effect of stimulus length on PLE and acceptability ratings of electrolaryngeal speech.

4.1.3 Chapter 4: Everyday listener impressions of perceived effort when listening to disordered speech

If PLE is shown to provide unique information at similar levels of intelligibility and acceptability, it will then be important to determine how best to measure it. As Table 1.1 shows, many types of scales have been used to quantify PLE, however, the best way to scale it is not clear. If PLE is to be used as a clinical outcome for disordered speech and voice, psychometrically valid ways of obtaining and analyzing ratings must be developed. Moreover, maximizing the ecological validity of PLE ratings applied to disordered speech or voice requires the use of untrained, unfamiliar listeners. Research addressing how these listeners interpret the task of rating their own effort is reported in this chapter.

Finally, given that internal listener characteristics affect measures of LE, any measures are likely to be affected by individual listeners' working memory capacity, receptive vocabulary, age and motivation. Future research should address how these variables interact with ratings of PLE, with the goal of developing strategies for listeners who have particular difficulty with disordered speech or voice.

References

Amichetti, N. M., Stanley, R. S., White, A. G., & Wingfield, A. (2013). Monitoring the capacity of working memory: Executive control and effects of listening effort. *Memory & Cognition*, *41*(6), 839–849. doi:10.3758/s13421-013-0302-0

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417–423.

Bennett, S., & Weinberg, B. (1973). Acceptability ratings of normal, esophageal, and artificial larynx speech. *Journal of Speech and Hearing Research*, *16*, 608–615.

Bernarding, C., Strauss, D. J., Hannemann, R., & Corona-Strauss, F. I. (2012). Quantification of listening effort correlates in the oscillatory EEG activity: A feasibility study. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 4615 –4618). doi:10.1109/EMBC.2012.6346995

Bernarding, C., Strauss, D. J., Hannemann, R., Seidler, H., & Corona-Strauss, F. I. (2013). Neural correlates of listening effort related factors: Influence of age and hearing impairment. *Brain Research Bulletin*, *91*, 21–30. doi:10.1016/j.brainresbull.2012.11.005

Bernarding, C., Strauss, D. J., Latzel, M., Hannemann, R., Chalupper, J., & Corona-Strauss, F. I. (2011). Simulations of hearing loss and hearing aid: Effects on electrophysiological correlates of listening effort. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC* (pp. 2319–2322). IEEE. doi:10.1109/IEMBS.2011.6090649

Beukelman, D., Childes, J., Carrell, T., Funk, T., Ball, L. J., & Pattee, G. L. (2011). Perceived attention allocation of listeners who transcribe the speech of speakers with amyotrophic lateral sclerosis. *Speech Communication*, *53*(6), 801–806.

Beukelman, D., Gillespie, L., Fager, S., & Ullman, C. (2014). Perceived attention allocation of listeners who transcribe the speech of dysarthric speakers with traumatic brain injury. *Journal of Medical Speech-Language Pathology*, *21*(3), 261–266.

Brons, I., Houben, R., & Dreschler, W. A. (2013). Perceptual effects of noise reduction with respect to personal preference, speech intelligibility, and listening effort. *Ear and Hearing*, *34*(1), 29–41. doi:10.1097/AUD.0b013e31825f299f

Brungart, D., Iyer, N., Thompson, E., Simpson, B. D., Gordon-Salant, S., Shurman, J., … Grant, K. W. (2013). Interactions between listening effort and masker type on the energetic and informational masking of speech stimuli. *Journal of the Acoustical Society of America*, *133*(5), 3435. doi:10.1121/1.4806059

Bunton, K., Kent, R. D., Duffy, J. R., Rosenbek, J. C., & Kent, J. F. (2007). Listener agreement for auditory-perceptual ratings of dysarthria. *Journal of Speech, Language & Hearing Research*, *50*, 1481–1495.

Carding, P. N., Wilson, J. A., MacKenzie, K., & Deary, I. J. (2009). Measuring voice outcomes: State of the science review. *Journal of Laryngology and Otology*, *123*(8), 823–829.

Choi, S., Lotto, A., Lewis, D., Hoover, B., & Stelmachowicz, P. (2008). Attentional modulation of word recognition by children in a dual-task paradigm. *Journal of Speech, Language & Hearing Research*, *51*(4), 1042–1054.

Cote-Reschny, K., & Hodge, M. (2010). Listener effort and response time when transcribing words spoken by children with dysarthria. *Journal of Medical Speech-Language Pathology*, *18*(4), 24–34.

Cullinan, W., Prather, E., & Williams, D. (1963). Comparison of procedures for scaling severity of stuttering. *Journal of Speech & Hearing Research*, *6*, 187–94.

Dagenais, P., Brown, G., & Moore, R. (2006). Speech rate effects upon intelligibility and acceptability of dysarthric speech. *Clinical Linguistics & Phonetics*, *20*(2/3), 141–148.

Desjardins, J. L., & Doherty, K. A. (2013). Age-related changes in listening effort for various types of masker noises. *Ear and Hearing*, *34*(3), 261–272. doi:10.1097/AUD.0b013e31826d0ba4

Donovan, N. J., Kendall, D. L., Young, M. E., & Rosenbek, J. C. (2008). The Communicative Effectiveness Survey: Preliminary evidence of construct validity. *American Journal of Speech-Language Pathology*, *17*(4), 335–347. doi>10.1044/1058-0360(2008/07-0010)</p>

Downs, D. W. (1982). Effects of hearing aid use on speech discrimination and listening effort. *Journal of Speech & Hearing Disorders*, *47*(2), 189–193.

Downs, D. W., & Crum, M. A. (1978). Processing demands during auditory learning under degraded listening conditions. *Journal of Speech and Hearing Research, 21*(4), 702–714.

Eadie, T. (2007). Application of the ICF in communication after total laryngectomy. *Seminars in Speech & Language*, *28*(4), 291–300.

Fairbanks, G. (1960). *Voice and articulation drillbook* (2nd ed.). New York: Harper.

Feuerstein, J. (1992). Monaural versus binaural hearing: ease of listening, word recognition, and attentional effort. *Ear and Hearing*, *13*(2), 80–86.

Fisk, A. D., Derrick, W. L., & Schneider, W. (1986). A methodological assessment and evaluation of dual-task paradigms. *Current Psychology*, *5*(4), 315–327. doi:10.1007/BF02686599

Fraser, S., Gagne, J.-P., Alepins, M., & Dubois, P. (2010). Evaluating the effort expended to understand speech in noise using a dual-task paradigm: The effects of providing visual speech cues. *Journal of Speech, Language & Hearing Research*, *53*(1), 18–33.

George, E. L. J., Zekveld, A. A., Kramer, S. E., Goverts, S. T., Festen, J. M., & Houtgast, T. (2007). Auditory and nonauditory factors affecting speech reception in noise by older listeners. *Journal of the Acoustical Society of America*, *121*(4), 2362–2375. doi:10.1121/1.2642072

Gosselin, P., & Gagne, J.-P. (2011). Older adults expend more listening effort than young adults recognizing speech in noise. *Journal of Speech, Language, and Hearing Research, 54*(3), 944–958.

Gosselin P.A, & Gagne J.-P. (2010). Use of a dual-task paradigm to measure listening effort. *Canadian Journal of Speech-Language Pathology and Audiology*, *34*(1), 43–51.

Gustafson, S., McCreery, R., Hoover, B., Kopun, J. G., & Stelmachowicz, P. (2014). Listening effort and perceived clarity for normal-hearing children with the use of digital noise reduction. *Ear and Hearing*. doi:10.1097/01.aud.0000440715.85844.b8

Hallgren, M., Larsby, B., Lyxell, B., & Arlinger, S. (2005). Speech understanding in quiet and noise, with and without hearing aids. *International Journal of Audiology*, *44*(10), 574–583.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA TLX (task load index): Results of empirical and theoretical research. *Human Mental Workload*, *1*, 139–183.

Hicks, C. B., & Tharpe, A. M. (2002). Listening effort and fatigue in school-age children with and without hearing loss. *Journal of Speech, Language, and Hearing Research*, *45*, 573–584.

Houben, R., van Doorn-Bierman, M., & Dreschler, W. A. (2013). Using response time to speech as a measure for listening effort. *International Journal of Audiology*, *52*(11), 753–761. doi:10.3109/14992027.2013.832415

Kahneman, D. (1973). *Attention and effort.* Englewood Cliffs, N.J.: Prentice-Hall.

Klasner, E., & Yorkston, K. (2005). Speech intelligibility in ALS and HD dysarthria: The everyday listener's perspective. *Journal of Medical Speech-Language Pathology*, *13*(2), 127–140.

Klasner, E., Yorkston, K. M., & Lillvik, M. (2002). Everyday listeners' perspective on Amyotrophic Lateral Sclerosis and Huntington disease dysarthria: barriers and strategies in understanding distorted speech samples. *Journal of Medical Speech-Language Pathology*, *10*, 293–298.

Koelewijn, T., Zekveld, A. A., Festen, J. M., & Kramer, S. E. (2012). Pupil dilation uncovers extra listening effort in the presence of a single-talker masker. *Ear and Hearing*, *33*, 291–300. doi:10.1097/AUD.0b013e3182310019

Koelewijn, T., Zekveld, A. A., Festen, J. M., Rönnberg, J., & Kramer, S. E. (2012). Processing load induced by informational masking is related to linguistic abilities. *International Journal of Otolaryngology*, *2012*, Article ID 865731. doi:10.1155/2012/865731

Kramer, S. E., Kapteyn, T. S., Festen, J. M., & Kuik, D. J. (1997). Assessing aspects of auditory handicap by means of pupil dilatation. *Audiology*, *36*(3), 155–64.

Larsby, B., Hallgren, M., Lyxell, B., & Arlinger, S. (2005). Cognitive performance and perceived effort in speech processing tasks: Effects of different noise backgrounds in normal-hearing and hearing-impaired subjects. *International Journal of Audiology*, *44*(3), 131–143.

Law, I., Ma, E. P.-M., & Yiu, E. M.-L. (2009). Speech intelligibility, acceptability, and communication-related quality of life in Chinese alaryngeal speakers. *Archives of Otolaryngology - Head and Neck Surgery*, *135*(7), 704–711.

Lindblom, B. (1990). On the communication process: Speaker-listener interaction and the development of speech. *Augmentative & Alternative Communication*, *6*(4), 220–230.

Mackersie, C. L., & Cones, H. (2011). Subjective and psychophysiological indexes of listening effort in a competing-talker task. *Journal of the American Academy of Audiology*, *22*(2), 113–122. doi:10.3766/jaaa.22.2.6

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes*, *27*(7-8), 953–978. doi:10.1080/01690965.2012.705006

McAuliffe, M. J., Wilding, P. J., Rickard, N. A., & O'Beirne, G. A. (2012). Effect of speaker age on speech recognition and perceived listening effort in older adults with hearing loss. *Journal of Speech, Language, and Hearing Research*, *55*, 838–847. doi:10.1044/1092-4388(2011/11-0101)

McGarrigle, R., Munro, K. J., Dawes, P., Stewart, A. J., Moore, D. R., Barry, J. G., & Amitay, S. (2014). Listening effort and fatigue: What exactly are we measuring? A British Society of Audiology Cognition in Hearing Special Interest Group "white paper." *International Journal of Audiology*, 1–13. doi:10.3109/14992027.2014.890296

Munro, M. J., & Derwing, T. M. (1995). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language & Speech*, *38*(3), 289–306.

Nagle, K., & Eadie, T. (2012). *Listener effort as an outcome measure for adductor spasmodic dysphonia*. Poster presented at the ASHA Convention, Atlanta, GA.

Nagle, K. F., & Eadie, T. L. (2012). Listener effort for highly intelligible tracheoesophageal speech. *Journal of Communication Disorders*, *45*(3), 235–245. doi:10.1016/j.jcomdis.2012.01.001

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204–238. doi:10.1016/S0010-0285(03)00006-9

O'Halloran, R., Hickson, L., & Worrall, L. (2008). Environmental factors that influence communication between people with communication disability and their healthcare providers in hospital: A review of the literature within the International Classification of Functioning, Disability and Health (ICF) framework. *International Journal of Language & Communication Disorders*, *43*(6), 601–32.

Oelke, M., Sachet, L., Nagle, K., Bislick, L., Brookshire, C., & Kendall, D. (submitted). Can intensive phonomotor treatment modify accent? A phase I study. *Clinical Linguistics & Phonetics*.

Panico, J., & Healey, E. (2009). Influence of text type, topic familiarity, and stuttering frequency on listener recall, comprehension, and mental effort. *Journal of Speech, Language & Hearing Research*, *52*(2), 534–546. doi:10.1044/1092-4388(2008/07-0238)

Papesh, M. H., Goldinger, S. D., & Hout, M. C. (2012). Memory strength and specificity revealed by pupillometry. *International Journal of Psychophysiology*, *83*(1), 56–64. doi:10.1016/j.ijpsycho.2011.10.002

Pichora-Fuller, M. K., Schneider, B. A., & Daneman, M. (1995). How young and old adults listen to and remember speech in noise. *Journal of the Acoustical Society of America*, *97*(1), 593–608. doi:10.1121/1.412282

Picou, E. M., & Ricketts, T. A. (2014). Increasing motivation changes subjective reports of listening effort and choice of coping strategy. *International Journal of Audiology*, *0*, 1–9. doi:10.3109/14992027.2014.880814

Picou, E., Ricketts, T., & Hornsby, B. (2011). Visual cues and listening effort: individual variability. *Journal of Speech, Language, and Hearing Research*, *54*(5), 1416–1430.

Picou, E., Ricketts, T., & Hornsby, B. W. (2013). How hearing aids, background noise, and visual cues influence objective listening effort. *Ear & Hearing*, *34*(5), e52–e64. doi:10.1097/AUD.0b013e31827f0431

Rakerd B, Seitz PF, & Whearty M. (1996). Assessing the cognitive demands of speech listening for people with hearing losses. *Ear and Hearing*, *17*(2), 97–106.

Rönnberg, J., Lunner, T., Zekveld, A., Sörqvist, P., Lyxell, B., Dahlström, Ö., … Rudner, M. (2013). The Ease of Language Understanding (ELU) model: theoretical, empirical, and clinical advances. *Frontiers in Systems Neuroscience*, *7*, 1–17. doi:10.3389/fnsys.2013.00031

Rönnberg, J., Rudner, M., Foo, C., & Lunner, T. (2008). Cognition counts: a working memory system for ease of language understanding (ELU). *International Journal of Audiology*, *47*(S2), S99–S105.

Rudner, M., Lunner, T., Behrens, T., Thorén, E. S., & Rönnberg, J. (2012). Working memory capacity may influence perceived effort during aided speech recognition in noise. *Journal of the American Academy of Audiology*, *23*(8), 577–589. doi:10.3766/jaaa.23.7.7

Sussman, J. E., & Tjaden, K. (2012). Perceptual measures of speech from individuals with Parkinson's disease and multiple sclerosis: Intelligibility and beyond. *Journal of Speech, Language & Hearing Research*, *55*(4), 1208–1219. doi:10.1044/1092-4388(2011/11-0048)

Tamati, T. N., Gilbert, J. L., & Pisoni, D. B. (2013). Some factors underlying individual differences in speech recognition on PRESTO: A first report. *Journal of the American Academy of Audiology*, *24*(7), 616–634. doi:10.3766/jaaa.24.7.10

Whitehill, T., & Wong, C. (2006). Contributing factors to listener effort for dysarthric speech. *Journal of Medical Speech-Language Pathology*, *14*, 335–342.

Winn, M., & Edwards, J. R. (2013). The impact of spectral resolution on listening effort revealed by pupil dilation. *The Journal of the Acoustical Society of America*, *134*(5), 4233. doi:10.1121/1.4831554

Winn, M., Litovsky, R. Y., & Edwards, J. R. (2014). Measurement of spectral resolution and listening effort in people with cochlear implants. *The Journal of the Acoustical Society of America*, *135*(4), 2390. doi:10.1121/1.4877906

World Health Organization. (2002). Towards a common language for functioning, disability and health: ICF. World Health Organization, Geneva. Retrieved from http://www.who.int/classifications/icf/en/

Yorkston, K. M., Klasner, E. R., & Swanson, K. M. (2001). Communication in context: A qualitative study of the experiences of individuals with multiple sclerosis. *American Journal of Speech-Language Pathology*, *10*(2), 126–137. doi:10.1044/1058-0360(2001/013)

Zekveld, A. A., Festen, J. M., & Kramer, S. E. (2013). Task difficulty differentially affects two measures of processing load: The pupil response during sentence processing and delayed cued recall of the sentences. *Journal of Speech, Language, and Hearing Research*, *56*(4), 1156–1165. doi:10.1044/1092-4388(2012/12-0058)

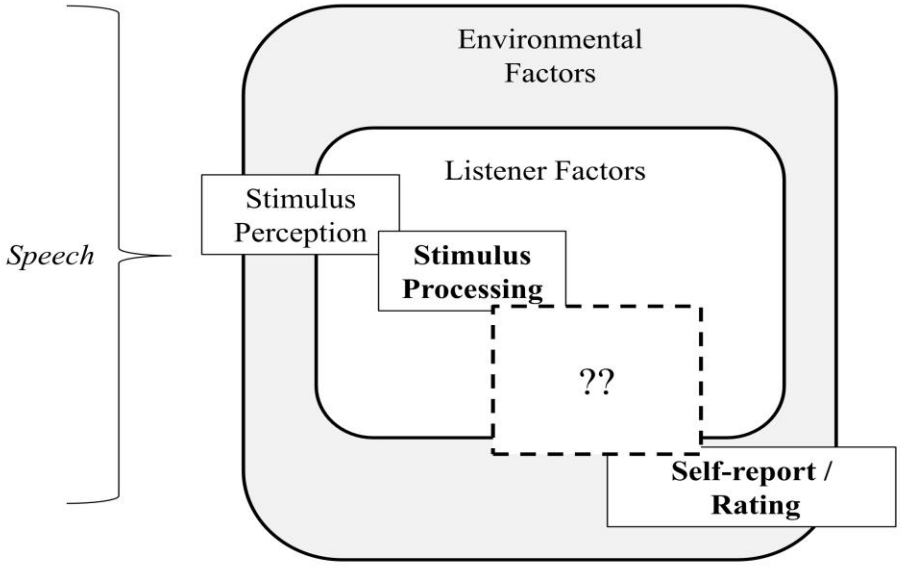
Zekveld, A. A., & Kramer, S. E. (2014). Cognitive processing load across a wide range of listening conditions: Insights from pupillometry. *Psychophysiology*, *51*(3), 277–284. doi:10.1111/psyp.12151

Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2011). Cognitive load during speech perception in noise: The influence of age, hearing loss, and cognition on the pupil response. *Ear and Hearing*, *32*(4), 498–510.

Zekveld, A., Kramer, S. E., & Festen, J. M. (2010). Pupil response as an indication of effortful listening: The influence of sentence intelligibility. *Ear and Hearing*, *31*(4), 480–490.

Appendix

Proposed model showing objective and subjective measures of listening effort



Listening effort (LE), represented here as **stimulus processing**, has been measured using EEG, fMRI and pupillometry in adverse listening conditions. It has also been measured by having people perform two simultaneous tasks and measuring the decrease in performance or processing speed on the secondary task (e.g., Fraser, Gagne, Alepins & Dubois, 2010; Gosselin & Gagné, 2011). These are direct measures of processing, but they do not align with the third way of measuring LE – **self-report** of perceived listener effort (PLE). At some point, maybe while still processing the stimulus, the listener may be making a judgment about it. The process of rating stimuli is not directly measurable; data obtained this way are subjective and possibly influenced by factors not included in more direct measures. The potential factors affecting self-reports and ratings are represented in this model using question marks.

Chapter 2
Perceived listener effort
for highly intelligible tracheoesophageal speech

Abstract

The purpose of this study was to determine whether: (a) inexperienced listeners can reliably judge listener effort and (b) whether listener effort provides unique information beyond speech intelligibility or acceptability in tracheoesophageal speech. Twenty inexperienced listeners made judgments of speech acceptability and amount of effort required to listen to 14 male tracheoesophageal speakers using a paired comparison paradigm. Intelligibility was controlled to limit the analysis to the relationship between ratings of listener effort and speech acceptability. Results showed that as a group, inexperienced listeners reliably rated both speech acceptability and listener effort. In addition, ratings of speech acceptability and listener effort were strongly correlated ($r > .99$); however, there was evidence that some individual listeners assigned different ratings for each dimension for the same speech samples. Results have important implications for communication success for tracheoesophageal speakers.

1. Introduction

Tracheoesophageal (TE) speech is an increasingly used method of voice restoration after total laryngectomy (Iverson-Thoburn & Haydon, 2000; Singer & Blom, 1980). The TE puncture procedure involves creating a fistula between the trachea and esophagus in order to link the lungs and reconstructed pharyngoesophageal (PE) segment; a one-way TE prosthesis is then placed in the puncture. When the tracheostoma is occluded on exhalation, pulmonary air is shunted through the prosthesis into the esophageal reservoir, setting the PE segment into vibration, and thereby creating the alaryngeal voice source for TE speech.

When compared to other alaryngeal speech methods (i.e., electrolaryngeal, esophageal), TE speech is consistently judged as most "preferred" and "natural" by listeners (Robbins, Fisher, Blom & Singer, 1984; Pindzola & Cain, 1988; Trudeau & Qi, 1990). However, TE speech is still described as rough, breathy or low in pitch, and is noticeably different from laryngeal speech and voice (Finizia, Dotevall, Lindström, & Lindstrom, 1998). The effects of this difference, along with the "effort" exerted by communication partners in listening to TE speech, remain a poorly understood but socially important outcome in this population.

1.1 Measuring outcomes in TE speech

A comprehensive approach to outcomes measurement is important after total laryngectomy. For example, patient-reported outcomes (e.g., health- or voice-related quality of life measures) are important indicators of success post-laryngectomy (Eadie, 2003), and complement more traditional measurements of the speech signal, such as evaluation of speech intelligibility, acoustic parameters of the speech signal, and auditory-perceptual judgments of speech and voice quality. However, because individual measures do not always directly relate to one another, a multidimensional approach to evaluation is ideal (Eadie, 2007).

Although speech intelligibility scores may provide an objective measure of speech production, they do not provide much information about the "differentness" of TE speech compared to laryngeal speech, partly because of a ceiling effect. For example, even a TE speaker who is 100% intelligible sounds noticeably different from a laryngeal speaker, and may require more effort on the part of the listener to understand. Highly intelligible TE speech has often been identified by listeners as less acceptable (Eadie & Doyle, 2005; Finizia et al., 1998) and less natural (Eadie & Doyle, 2002) than laryngeal speech. Simply put, intelligibility measures alone may not be sensitive to capturing qualities of TE speech that differentiate the performance of TE speakers. The challenge is in finding valid measures that reliably distinguish between highly intelligible TE and laryngeal speech, and among TE speech samples of equal intelligibility, in the presence of a perceptually obvious difference.

TE speech has historically been measured in terms of the deviation of the new voice from listener expectations of a typical voice. For this reason, multidimensional or global aspects of speech are usually measured for TE speech (Eadie & Doyle, 2002; 2005; Finizia et al., 1998; Pindzola & Cain, 1988; Trudeau, 1987). In general, results have shown that regardless of perceived communicative "excellence," TE speakers are judged as having voices that are less acceptable and poorer in voice quality than normal speakers (Bennett & Weinberg, 1973; Finizia et al., 1998; Pindzola & Cain, 1988).

When listeners rate the quality of a speech sample, they quantify its severity using some type of rating scale (Eadie & Doyle, 2002; 2004). Because of the transient nature of the speech signal, listeners sometimes have difficulty maintaining their internal standards (i.e., their percept of the dimension being judged) when making these judgments (Kreiman, Gerratt, Precoda & Burke, 1992). Reliability for these methods can therefore be quite variable for both intra- and

interrater comparisons. An alternative method for judging speech samples is to present stimuli in pairs to listeners (i.e., paired comparisons), and to have each listener make judgments about which stimulus best exemplifies the dimension. Comparing each stimulus with every other stimulus results in rank ordering among the speech samples. Reliability for paired comparisons is often stronger than traditional rating scales because listeners compare only one stimulus to one other, and are not asked to quantify how much a given stimulus demonstrates an attribute (Eadie, Doyle, Hansen & Beaudin, 2008; Meltzner & Hillman, 2005). The paired comparison method may circumvent some of the inherent difficulties and variability with other types of scales, although its use is often limited to research applications because of the time required to create stimulus pairs, and because there is no set of established external referents with which to compare alaryngeal samples.

While the various dimensions of voice quality are naturally relevant to outcomes, there is an additional factor that is critical to successful communication. As Kreiman and Gerratt (1996) pointed out in their discussion of multidimensional scaling, "[voice] quality cannot be treated solely as an attribute of voices" (p.1793). Sound quality, as described by the American National Standards Institute (ANSI), is ''that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar'' (ANSI Standard S1.1.12.9, p. 45, 1960). That is, the listener is inescapably part of the process in determining the ultimate success of the speaker. The acoustic and perceptual characteristics of the speech signal do not necessarily reflect the processing burden on the listener; for a complete picture we must examine the interactions among the signal, the task and the listener (Kreiman et al., 1993).

1.2 Listener burden

Investigation of the effect of listener burden on communication has been quite limited in the field of communication disorders, beyond research from the perspective of the hearing impaired (e.g., Anderson Gosselin, & Gagné, 2011; Zekveld, Kramer, & Festen, 2010). Given the interactions of task, signal and listener, however, perception of speech acceptability (or naturalness, severity, etc.) could be altogether different from the amount of effort required by the listener. For example, a very rough speech sample could easily be judged as highly unacceptable by a familiar listener, while requiring little effort to understand (i.e., low speech acceptability with unexpectedly low effort). In other words, individual listener effort may depend on features such as familiarity with the particular speaker, with a specific population of speakers, or with the specific type of speech produced by the speaker (e.g., disordered, accented, particularly hurried).

In examining the dimension of listener burden, the focus is shifted toward the listener (i.e., away from the signal itself), which obliges the listener to think about his or her own reaction to the speech. This focus on the perceived cognitive resources required to process the speech signal may reveal meaningful differences between the listener's impression of the perceptual qualities of the signal and the amount of effort required to process it (Beukelman, Childes, Carrell, Funk, Ball, & Pattee, 2011; Evitts & Searl, 2006).

Research on auditory-perceptual ratings of speech disorders has provided some evidence of listener burden as a unique construct (Beukelman et al., 2011; Healey, Gabel, Daniels & Kawai, 2007; Whitehill & Wong, 2006). In the fluency literature, the social validity of treatment is sometimes measured in terms of "listener comfort," although this term is usually not specifically defined (Evans, Healey, Kawai & Rowland, 2008; O'Brien, Packman, Onslow, Cream, O'Brian & Bastock, 2003). Instead, listeners in fluency studies are often asked to

indicate their level of agreement with such statements as "I felt comfortable listening to this boy" using an interval scale (i.e., 1 to 5 point scale; Evans et al., 2008). Listeners in some studies describe how comfortable they were listening to a speaker in response to an open-ended question (Healey et al., 2007). Finally, other studies have used the term listener comfort to "reflect[ing] feelings about the way the person speaks, not what the person is saying or how their personality affected" them (O'Brian et al., 2003). O'Brien and colleagues (2003) examined the difference between a group of inexperienced listeners using a scale of listener comfort and a similar group rating speech naturalness for pre- and post-treatment samples of dysfluent speech. Although intrarater reliability was essentially equal for the two scales (r = .79, r = .78, respectively), mean ratings of listener comfort were much less reliable (ICC = .50) than those of speech naturalness (ICC = .71). Nevertheless, post-treatment ratings of listener comfort were significantly higher than pre-treatment ratings, indicating some level of clinical usefulness. Finally, the similarity of pre-treatment ratings of listener comfort and speech naturalness compared to the significant difference between post-treatment ratings for the dimensions suggested that the concept of listener comfort appeared to capture a dimension other than speech naturalness.

One other published study has investigated the construct of listener comfort in individuals with adductor spasmodic dysphonia (Eadie, Nicolici, Baylor, Almand, Waugh & Maronian, 2007). The authors found that inexperienced listeners judged listener comfort reliably (i.e., intrarater reliability correlation coefficient =.89; interrater reliability alpha coefficient = 0.98). Together, the results of the O'Brian et al. (2003) and Eadie et al. (2007) studies suggest that listener comfort is a viable construct that needs future examination as an outcome measure, but that it may be difficult to use reliably for some types of speech and voice disorders.

In contrast to listener comfort, the term "listener effort" (hereinafter "perceived listener effort," or PLE) has been used in dysarthria research to address the increased burden (cognitive processing load) placed on the listener by disordered speech (Klasner & Yorkston, 2005). Whitehill and Wong (2006) studied the relationship between PLE and intelligibility in dysarthric speakers. In addition to intelligibility, they asked listeners to judge the "amount of effort required" to listen to the samples. Although a strong correlation between intelligibility and PLE was found (r_s = -.95), there were three speech samples with equally high intelligibility which were also judged as requiring "high effort." This finding of discontinuity between intelligibility and PLE for some speakers supports the idea that PLE may capture factors beyond intelligibility, and is bolstered by a recent study of "attention allocation" (Beukelman et al., 2011).

Beukelman and colleagues (2011) examined the relationship of attention allocation, or the amount of work a listener expends in having a conversation with a person with disordered speech, to speech intelligibility in speakers with amyotrophic lateral sclerosis (ALS; Beukelman et al., 2011). Mean scores from the Sentence Intelligibility Test (SIT) for their 32 speakers ranged in intelligibility from 3.6% to 100%, and scores from the five listeners spanned the range of 1.7 to 6.73 on a 7-point Likert scale for self-perception of attention allocation. There was a predictable relationship between attention ratings and intelligibility, with a correlation of -.89, but the highest ratings of attention allocation were given to speech samples that were 75% to 80% intelligible. In fact, several of the speakers in this study whose mean SIT scores were 90% or greater received attention allocation scores in the middle of the scale (i.e., 4 points on the 7-point scale). These findings suggest that some dimension of listener burden, whether called effort, attention, or some other name, encompasses paralinguistic parameters beyond intelligibility, and may include listener factors independent of the speech signal itself.

Klasner and Yorkston (2005) investigated PLE in a qualitative study on the barriers to communication and strategies used by listeners to understand dysarthric speech. Statements elicited from their listeners are similar to what might be expected from listeners judging any kind of distorted speech:

"It was hard to listen to this sentence."

"I got distracted by the way the speech sounded."

"I had to be prepared to hear distorted speech."

"I had to completely attend to the sentence to understand it."

"I had to concentrate on understanding the sentence." (p. 134)

These statements make it clear that while listeners may eventually interpret 100% of the words spoken in a speech sample, they have to prepare themselves to do so for some types of speech. If speech samples of equal intelligibility are not equally natural (or pleasant, or acceptable), it is reasonable to assume that they may not require equal effort on the part of the listener; that is, while a listener may eventually understand a TE speaker completely, the effort expended to do so may be significant (and significantly different from that required to listen to a laryngeal speaker). It is logical to conclude that listeners may decline to initiate or maintain communication with a speaker who imposes an increased burden on them, making PLE an important construct to consider.

Despite the difficulty of finding suitable objective measures of alaryngeal speech, a single published study has instrumentally investigated listener processing demands for decoding it (Evitts & Searl, 2006). Specifically, Evitts and Searl (2006) measured reaction times in naïve listeners making judgments of laryngeal, synthetic, and alaryngeal methods of speech. One highly intelligible, representative speaker was chosen for each method (i.e., TE, esophageal,

electrolaryngeal and laryngeal speech), and additional samples were synthesized for comparison. Listeners indicated whether the single-word speech sample was the same or different as an orthographic stimulus on a computer screen. Reaction time ratios were calculated to compare the five types of speech. Results indicated that cognitive processing loads for single word stimuli in TE speech were comparable to those for normal speech. Caution is warranted in generalizing these findings, however; only single-word stimuli and only one proficient speaker per condition were used to examine these effects. Additionally, the effect of using different modalities in this study (both auditory and written/visual) is not known. Finally, it is unclear how listener processing demands measured by reaction times differs from scaled measures of PLE. As a consequence, an investigation of this concept using perceptual measures appears warranted.

1.3 Experimental questions

In summary, although several dimensions of TE speech have been examined as valid post-laryngectomy outcomes, empirical investigation of perceived listener effort in alaryngeal speakers has been limited. This dimension is important to investigate because listener burden may relate to the willingness of a communication partner to engage the speaker. Although PLE is primarily a feature of the listener and appears to be different from other aspects of the speech signal such as speech intelligibility (Beukelman et al., 2011; Whitehill & Wong, 2006), it is unknown whether it can be differentiated from traditional measures of TE speech such as severity, naturalness, pleasantness, or acceptability. For example, Eadie and Doyle (2005) described “speech acceptability” as a dimension addressing both the listener’s burden and the consequent social impact of a distorted voice signal. To begin to test the utility of a construct involving any perceptual dimension, it is first necessary to determine whether listeners are able

to reliably judge the dimension using an appropriate scaling method. It is also necessary to determine whether the dimension is at least somewhat differentiated from existing standard measures. Consequently, this study was primarily designed to answer the following two questions:

1. Can inexperienced listeners reliably judge perceived listener effort in TE speech?
2. Is perceived listener effort a viable construct in TE speech? That is, does perceived listener effort provide unique information not captured by constructs such as speech acceptability or intelligibility?

2. Methods

2.1 Stimuli and preparation

Speech samples from 14 adult male, native English speakers were obtained from an archived database. Speakers were at least six months post-laryngectomy and used TE speech as their primary mode of communication. They ranged in age from 42 to 78 years (mean = 63 years). In order to control the effects of intelligibility on PLE and acceptability, all of the chosen speech samples were recordings of the second sentence of Fairbanks' Rainbow Passage (Fairbanks, 1960). Two experienced speech-language pathologists individually rated potential samples on a 5-point scale for PLE and speech acceptability to ensure the selected samples displayed a range of each dimension. To control possible effects of dialect on intelligibility, acceptability or PLE, only samples spoken in a Standard American English dialect were selected.

Speech samples were normalized for peak intensity and edited to create paired samples using acoustic software (Sony Soundforge 7.0). In the interest of presenting the samples as realistically as possible, the noise often associated with the onset of TE speech was not cut from the samples. Each speaker sample was paired with every other sample in A-B and B-A

conditions (n = 14 x 13 = 182) in a standard paired comparison paradigm; voices within a pair were separated by 0.5 seconds (Kreiman & Gerratt, 1996). Paired samples were then entered into a custom-made software program (Ruby on Rails) designed to randomize speaker pair presentation and to obtain listener responses on rating scales. Eighteen speaker pairs (10%) per dimension were randomly repeated to determine intrarater reliability, resulting in 200 judgments per listener for each dimension.

2.2 Listeners

Twenty adult native English speakers (12 female, 8 male) with no prior exposure to alaryngeal voice were recruited for this study. All were considered inexperienced listeners. Listeners ranged in age from 18 to 32 years (mean = 23.4 years). They reported no concerns about their hearing and passed hearing screening tests at 25 dB at the octave frequencies between 250-4000 Hz.

2.3 Procedures

Before any judgments were made, listeners were familiarized with the task and provided definitions of acceptability and PLE (see Appendix). For the purposes of this study, PLE was defined as "the amount of work needed to listen to a speaker" (Whitehill & Wong, 2006, p. 337). When rating speech acceptability, listeners were asked to "Give careful consideration to the attributes of pitch, rate, understandability, and voice quality. In other words, is the voice acceptable to listen to as a listener?"(Bennett & Weinberg, 1973, p. 610). Use of this definition of acceptability permitted comparison of these results with those in the previous alaryngeal literature (Eadie & Doyle, 2005; Finizia et al., 1998; Pindzola & Cain, 1988).

During each session, listeners sat in front of a computer screen, and heard stimuli over headphones (Samson Stereo Headphones, RH600) set to a comfortable volume. The stimuli, the

second sentence of the Rainbow Passage, were presented using the custom-made software program (Ruby on Rails). Each stimulus pair was presented only once per trial. Listeners controlled the rate of presentation of the sample pairs, but were unable to replay a stimulus.

Listeners were asked to judge which sample in each pair required *less* effort or was *more* acceptable, to keep the scale similar for both dimensions. Samples were judged using an undifferentiated 100 mm visual analog scale (VAS), marked at the end points (0 mm = speaker 1 is less effortful/more acceptable; 100 mm = speaker 2 is less effortful/more acceptable). A judgment in the middle of the line (at 50 mm, or "neutral") indicated that the speakers required equal amounts of effort or were equally acceptable. In this way, a confidence rating was built into the scale; the farther from midline, the more "preferred" the sample would be (Searl & Small, 2002).

Each listener judged all 14 speaker samples in both A-B and B-A pairings for each rating dimension. Listeners judged one dimension (PLE or acceptability) in the first rating session, with the order of stimuli and dimension counterbalanced across listeners. The second dimension was judged in a second session held at least one week (but no more than three weeks) later to control for learning effects. All recruitment methods and procedures were approved by the University of Washington Human Subjects Committee, and all listeners were paid for their participation.

2.4 Data analysis

Data are reported and analyzed in raw and converted form, based on individual listener data and group means for all listeners. Raw "discrete speaker ratings" were measured in millimeters from the far left point of the scale (at 0 mm) and converted to allow comparison of sample scores. Scores favoring Sample 1 (i.e., to the left of "neutral," [50 mm]) were subtracted from 100 for comparison with scores favoring Sample 2 (to the right of neutral). A sample with

a converted score of 100 was interpreted as "Definitely More Acceptable" or "Definitely Less Effort" than the other sample in the pair. Likewise, a sample with a converted score of 25 was interpreted as less acceptable or requiring more PLE than the other sample in the pair. Scores in the middle of the range (40 to 60 mm) were taken to indicate no preference of sample for the given dimension (Searl & Small, 2002). "Average speaker ratings" were established based on the mean discrete ratings for each speaker from all listeners (13 speaker pairs x 2 stimulus orders x 20 listeners = 520 judgments per speaker per dimension).

To answer the experimental question of whether inexperienced listeners can reliably judge PLE, reliability and variability coefficients were calculated. Reliability and variability were also determined for speech acceptability to ensure the representativeness of the listener group's use of the existing standard measure. Intrarater reliability was calculated for each dimension using the first and second ratings for repeated stimuli (n = 18) to derive Pearson product moment correlation coefficients for individual listeners. Interrater reliability was calculated by comparing each listener's ratings to each other listener's ratings and by comparing each listener's ratings to the group mean for each speaker. The relationships between individual listener ratings were examined using single-measures intraclass correlation coefficients (ICCs), and the relationships between individual listener ratings and group means were evaluated using average-measures ICCs (Shrout & Fleiss, 1979).

Interrater variability, a measure of the dispersion of scores around a mean value, was also established for each of the 182 sample pairs. Unlike measures of interrater agreement, this measure considers the variability of listener ratings without using an arbitrary cutoff point, such as "within 10 mm" (Chan & Yiu, 2002; Portney & Watkins, 2000).

The second experimental question, regarding the validity of PLE, was addressed by comparing mean ratings of each dimension for each speaker and by examining differences between individual listeners' ratings of the same samples for PLE and for speech acceptability. The relationship was determined using a Pearson's Correlation Coefficient. Matched pair *t* tests with Bonferroni corrections ($p < .0025$) were also calculated for individual listener data to determine the significance of individual listener differences in ratings of PLE and acceptability for the same sample pairs. Finally, ratings for the two dimensions were compared for each sample pair, based on whether they fell within the range of Speaker 1 (0-39 mm) or Speaker 2 (61-100 mm); neutral ratings were ignored. For example, if a rating fell in the range of Speaker 1 for acceptability, but in the range of Speaker 2 for PLE, this was interpreted as a meaningful difference in perception of these dimensions for that sample pair. Whereas the results of *t*-tests might reveal a systematic difference in ratings between dimensions, the analysis of differences per sample was meant to reveal larger differences between ratings of acceptability and PLE assigned by the same rater to the same speech sample.

3. Results

3.1 Discrete ratings

Raw scores were converted to allow comparison of ratings within and between listeners, and to compare similarity of ratings of PLE to speech acceptability. Using the rating scales provided, a lower score indicated more PLE, but less speech acceptability; this allowed comparison of mean converted discrete scores across dimensions. Most listeners used the entire range of the scale, from 0-100 mm, to rate each dimension. Average converted discrete ratings ranged from 24.53-78.79 for PLE and from 20.91-80.90 for speech acceptability, as shown in

Table 2.1. The order of mean ratings from lower to higher scores was consistent for the two dimensions, except for speakers 15 and 16 (which were reversed).

*Table 2.1***.** Mean ratings for all listeners for acceptability and listener effort on 100 mm. visual analog scale, in mm, arranged from lowest to highest acceptability score. Higher scores indicate greater acceptability or less effort. Italics indicate reversed order for PLE compared to acceptability

Speaker #	**Acceptability Mean (SD)**	**Perceived Listener Effort Mean (SD)**
11	20.91 (11.82)	24.53 (11.72)
19	31.70 (17.72)	32.36 (15.93)
22	32.19 (16.93)	34.52 (16.65)
17	34.99 (18.38)	36.65 (18.05
21	39.53 (19.77)	40.22 (18.85)
16	42.86 (19.97)	*41.26 (18.86)*
15	43.09 (19.92)	*41.10 (16.32)*
4	47.47 (21.30)	48.08 (18.11)
24	48.81 (18.53)	51.05 (20.39)
9	61.52 (20.27)	59.16 (18.36)
14	61.83 (19.63)	63.17 (16.01)
12	74.18 (18.02)	72.08 (15.72)
27	80.04 (13.52)	77.03 (14.63)
20	80.90 (14.91)	78.79 (13.97)

3.2 Intrarater reliability

Mean Pearson correlation coefficients for each listener indicated that individual measures of intrarater reliability ranged from r = .50 - .94 (mean r = .78, SD = 0.11) for PLE, and from r = .56 - .94 (mean r = .78, SD = 0.10) for speech acceptability for the 18 repeated sample pairs. As shown in Table 2.2, there were some large differences in reliability between the dimensions for some listeners, although the difference between group average reliability for acceptability and PLE was not significant [$t(19) = -0.15$; $p = .883$].

*Table 2.2***.** Intrarater reliability for speech acceptability and listener effort for individual listeners. Pearson's correlation coefficients indicating the correlation between repeated ratings within each dimension are displayed.

	Intrarater Reliability	
Listener #	**Acceptability *(r)***	**Listener Effort *(r)***
1	.669	.800
2	.927	.943
3	.762	.891
4	.659	.782
5	.854	.860
6	.886	.742
7	.918	.500
8	.810	.875
9	.728	.779
11	.837	.533
12	.899	.683
13	.812	.762
14	.945	.882
15	.649	.689
16	.781	.744
17	.766	.866
18	.719	.811
19	.719	.881
20	.718	.760
21	.562	.731
Mean (SD)	.775 (.114)	.781 (.105)

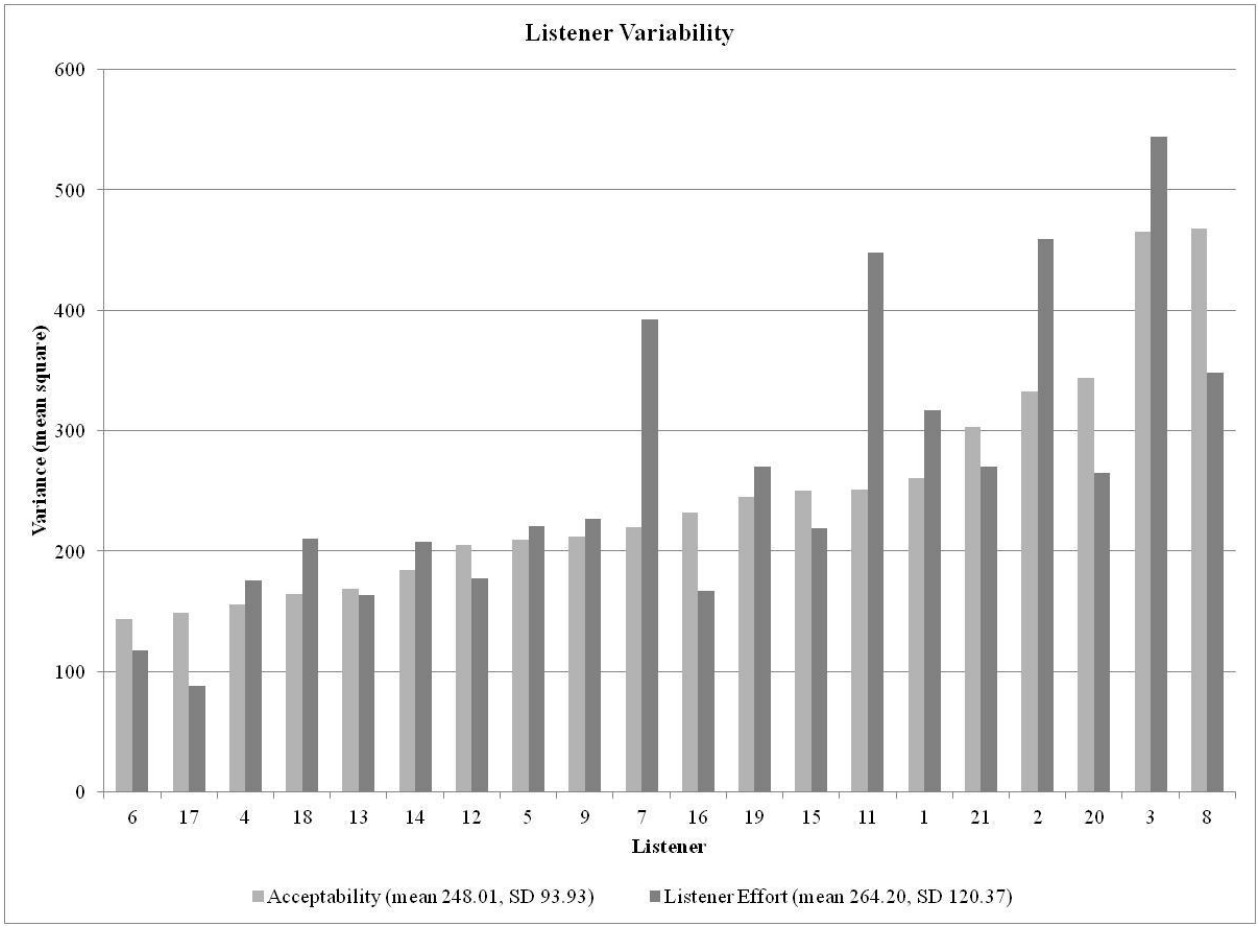
3.3 Interrater reliability

Single-measures ICCs represent the reliability of each listener compared to each other listener. Based on single-measures ICCs, interrater reliability was good for both PLE (r = .66) and for speech acceptability (r = .71; Portney & Watkins, 2000). Average-measures ICCs

represent the reliability of each listener compared to the mean for each speaker. Based on average-measures ICCs, interrater reliability was very strong for PLE (r = .98) and for speech acceptability (r = .98).

The sample variance for PLE was 264.20 (SD = 120.37), with a range of 87.99 to 544.15; for acceptability the sample variance was 248.01 (SD = 93.93), with a range of 143.19 to 468.09. The difference in variance between the two dimensions was not significant [$t(19) = -0.89$; p =.383]. The variability of individual listeners (arranged in increasing order of acceptability rating) is displayed graphically in Figure 2.1.

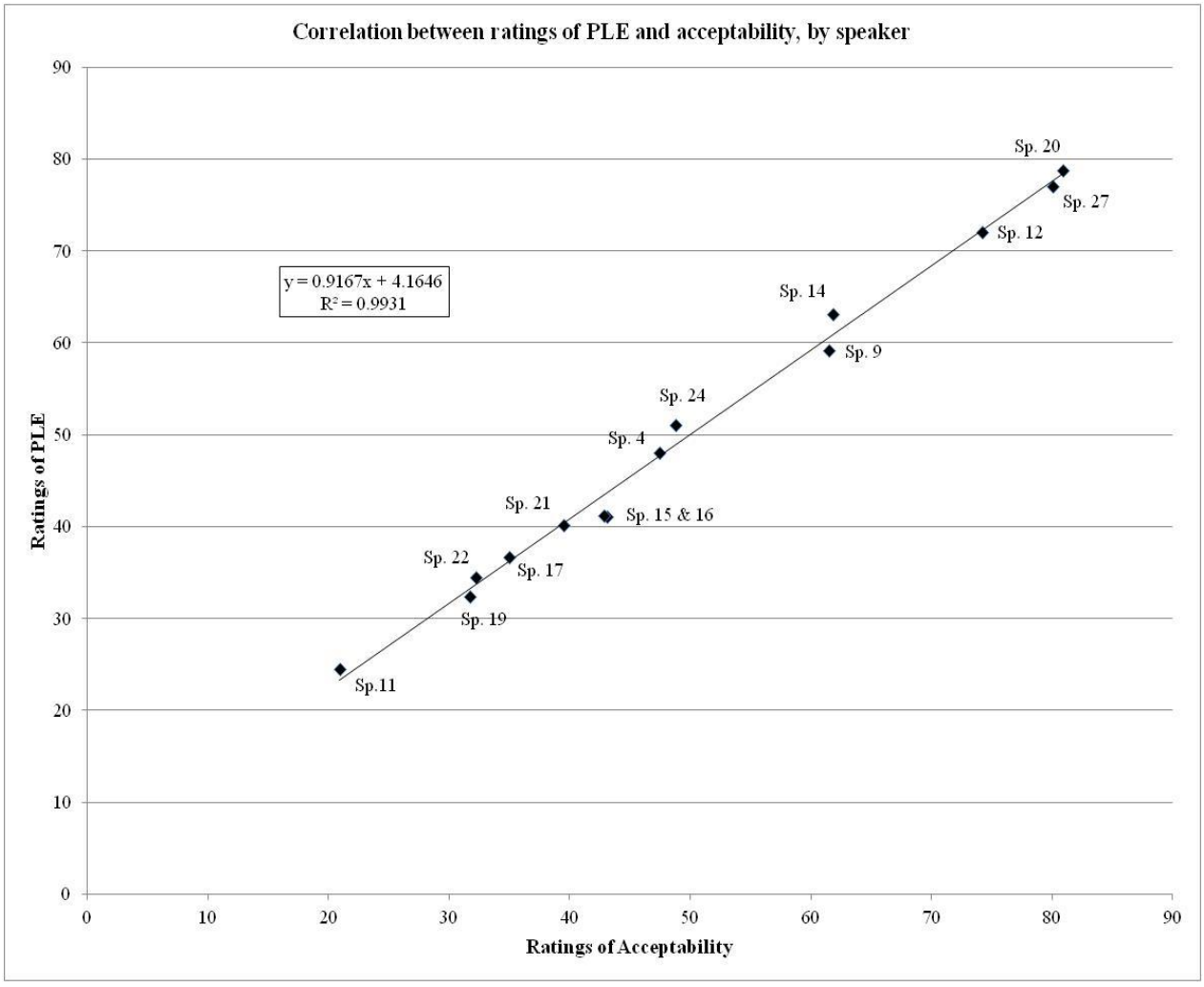
Figure 2.1. Interrater variability for the dimensions of speech acceptability and perceived listener effort, expressed in mean squares for each listener.



3.4 Relationship between perceived listener effort and speech acceptability

The relationship between the two dimensions was determined to establish whether ratings of PLE provided unique or additional information from that provided by ratings of speech acceptability. In addition to the almost identical order of scores shown in Table 2.1, the Pearson's correlation between mean ratings of the two dimensions by speaker was very strong (r = .99), and a linear relationship with tight distribution about the line of best fit was revealed (see Figure 2.2).

Figure 2.2. Correlation between mean discrete scores in mm (0-100) for perceived listener effort and acceptability, by speaker.



Pearson correlation coefficients were also calculated for individual listener ratings of all 200 sample pairs (182 original samples plus 18 repeated samples) to establish the extent to which individual listeners varied in their understanding of the two dimensions. Correlations between individual listener ratings of PLE and acceptability for the same samples ranged between r = .60 - .87 (mean r = .77, SD = 0.08).

To determine the significance of differences in each listener's discrete ratings for speech acceptability compared to PLE for the same sample, two-tailed, matched pair *t*-tests ($p < .05$) were also performed on the discrete data for the two dimensions. Because these data cannot be assumed to be independent, a Bonferroni adjustment to the alpha level of .05 was made to control for error (level of significance = .0025). Only two listeners' (1 and 16) ratings were found to be significantly different ($p = .001, .002$, respectively), despite the moderate to strong correlations ($r = .69, .76$) for their ratings of speech acceptability and PLE.

Comparison of individual ratings of each sample across dimensions indicated that 3% of total listener ratings differed depending on which dimension was being rated. Nearly all listeners (19/20) assigned ratings indicating a preference for a different speaker depending on the dimension for at least one speech sample (mean 3%, range 0-8% of ratings). No relationship was found between individual listener reliability and tendency to change speaker preference. In fact, Listener 18 (who did not change any speaker preference) and Listener 4 (who changed the most) were equally reliable across dimensions (see Table 2.2).

4. Discussion

This study had two purposes: to determine whether inexperienced listeners can reliably judge their own effort when listening to TE speech, and to establish whether these ratings provide differential information above and beyond speech acceptability and intelligibility. Results showed that as a group, inexperienced listeners reliably rated both speech acceptability and PLE.

Notably, a range of mean discrete ratings was assigned to both acceptability (20.91-80.90) and PLE (24.53-78.79), for samples of equally highly intelligible speech (i.e., near perfect). Given the wide range of scores exhibited across both acceptability and PLE in the

current study, the findings support the conclusion that these constructs are clearly perceptually different from intelligibility. These findings are consistent with earlier examinations of the acceptability of highly intelligible TE speech (Eadie & Doyle, 2005; Finizia et al., 1998). It may be that one or both of these constructs addresses the "differentness" alaryngeal speech from laryngeal speech, or one alaryngeal speaker from another.

The dimensions of speech acceptability and PLE were also found to be strongly correlated ($r > .99$), based on mean ratings for each speaker. When individual listener data were analyzed separately, however, a different pattern emerged. As might be expected, there was a wide range of reliability among listeners, but there was also a large difference in reliability between the dimensions of speech acceptability and PLE for some listeners. Though not significant, individual listeners tended to rate speech acceptability of TE speakers more reliably than PLE. These results suggest that the term or the concept of acceptability might have more perceptual reality for inexperienced listeners than a construct such as "perceived listener effort." Additionally, most listeners rated at least one sample pair differently for each dimension. These differences provide some initial evidence that individual listeners may use different strategies, or have different criteria for these two dimensions. These results have implications for measuring outcomes in alaryngeal speech.

4.1 Reliability of judgments

The successful use of rating scales depends on listeners judging the same sample in the same way for a given dimension each time they hear it. Overall, mean intrarater reliability for mean ratings of both PLE and speech acceptability in the current study ($r = .78$) was consistent with previous studies using the dimension of listener comfort (Eadie et al., 2007; O'Brien et al., 2003). Despite relatively strong overall within-listener reliability, however, there was a wide

range of reliability for ratings of the two dimensions for individual listeners. For example, the correlation between Listener 7's first and second ratings for speech acceptability was very strong (r = .92), but noticeably less for PLE (r = .50; see Table 2.2). This example suggests that while the theoretical "average" listener is capable of rating both dimensions with strong reliability, individual raters may need additional information in order to increase intrarater reliability.

In order to be clinically useful, rating scales must also be used similarly by different listeners. Typically, a listener's results are compared with those of an average listener. The correlations between each listener and the group mean in this study were very strong for both dimensions (average-listener ICCs r > .97). This indicates that on average, each listener's ratings were very similar to the group mean for each speaker sample. Interrater reliability for the current study is consistent with one previous study examining listener comfort and speaker effort in spasmodic dysphonia (Eadie et al., 2007). However, the results are considerably higher than the ICCs reported by O'Brian and colleagues (2003), in their examination of listener comfort (r = .50) and speech naturalness (r = .71), or the ratings of PLE (r_s = .67) for dysarthric speech (Whitehill & Wong, 2006). Several factors may account for this difference, including differences in population, linguistic level of speech samples, mode of presentation, and scale type. For example, O'Brian and colleagues (2003) presented 30-second samples of dysfluent conversation in video format, while the current study used alaryngeal samples of a read sentence presented in an auditory-only format. Additionally, the 9-point equal-appearing interval scale used by O'Brian and colleagues (2003) required listeners to consult their own internal referents in quantifying the attribute in question, whereas the paired comparison method used in the current study required only that listeners compare the first sample to the second. The paired comparison method may also promote increased reliability among judges (Maryn et al., 2009).

Since no listener is truly "average," each listener's ratings were also compared with those of every other listener. As may be expected, given the range of variability for each listener, this relationship was only moderate. The single-listener ICC for speech acceptability and PLE showed moderate correlation among listeners (r = .71 and r = .66, respectively). This moderate relationship was also observed in measures of variability, as shown in Figure 2.1. The mean variance of ratings by listener was similar for both dimensions.

Together, the results examining reliability showed that listeners appeared to be equally consistent using both dimensions of speech acceptability and PLE. The equivalently strong reliability for the two dimensions is interesting, as speech acceptability is a much more widely used and recognized term in the measurement of alaryngeal speech outcomes (Eadie & Doyle, 2005; Finizia et al., 1998; Pindzola & Cain, 1988; Trudeau, 1987). The ability of the average listener to rate these dimensions with equal consistency suggests that neither "acceptability" nor "perceived listener effort" is inherently a "better" descriptor of the experience of listening to an alaryngeal speech sample. Future research may determine that this similarity indicates that ratings of PLE add limited information to currently used outcome measures. However, to determine whether ratings of PLE add any independent information, it is first necessary to examine the relationship between these measures.

4.2 Relationship between perceived listener effort and speech acceptability

To determine whether ratings of PLE capture features not included in ratings of speech acceptability, correlations between ratings of each dimension were calculated by listener and by speaker. The very strong correlation between ratings of PLE and speech acceptability may provide evidence that they are expressing the same information (r = .99). In fact, Eadie and colleagues (2007) reported a very strong correlation between vocal effort and listener comfort

(Pearson's r = -.98) and between overall severity and listener comfort (r = .98) for samples of speech with adductor spasmodic dysphonia. Initial ratings of listener comfort and speech naturalness were also highly correlated (r = .96), although post-treatment ratings were not strongly related (r = .46) for dysfluent samples (O'Brian, et al., 2003). Finally, the correlation between PLE and intelligibility was strong (Spearman's r = -.95) for dysarthric samples (Whitehill & Wong, 2006).

Despite group data that reveal a strong correlation between the average listener's responses for the two dimensions, there are some reasons to believe that ratings of PLE may actually capture information not expressed in ratings of speech acceptability. First, the instructions given to listeners were quite different for each dimension. Listener effort was defined rather broadly as the amount of work needed to listen to the speaker. Acceptability, on the other hand, was presented in specific terms; listeners were asked to focus on "attributes of pitch, rate, understandability and voice quality" in making their overall acceptability judgments. At least in terms of the vocabulary used to describe them, these were different tasks.

Second, despite very strong overall group mean correlations, the relationships between individual listener ratings of PLE and acceptability for the same samples were noticeably different (ranging from r = .60 - .87, with mean r = .77). In fact, two listeners demonstrated a statistically significant difference in ratings of samples depending on the dimension. Reliability for both Listener 1 and Listener 16 was in the average range for both dimensions, and their individual ratings for speech acceptability were significantly different from those for PLE for the same samples.

Third, post hoc analysis of speaker preference indicated ratings of the two dimensions for the same samples sometimes differed enough to change speaker preference from Speaker 1 to

Speaker 2 and vice versa. This is evidence that individual listeners sometimes assigned different ratings for each dimension for the same speech sample. Listeners were using a continuum with one speaker at each endpoint, as opposed to rating the magnitude of a dimension for a single sample. Given that we have interpreted the "neutral" area as the range between 40-60 mm on the VAS (Searl & Small, 2002), differences this large led to almost a categorical change. For example, a rating of 39 for acceptability coupled with a rating of 80 for PLE for one sample pair suggests that for a particular listener, Speaker 1's sample was more acceptable, but Speaker 2's sample required less effort.

Finally, nearly all of the listeners changed speaker preference for at least one sample pair across dimensions, and no relationship was found between intrarater reliability and number of changed speaker preferences across dimensions. Differences in speaker preference combined with relatively robust intrarater reliability strongly suggest that most listeners made judgments based on different criteria for each dimension, and that they were consistent in their ratings within dimensions.

4.3 Future directions

To determine the perceptual basis of PLE and acceptability, and whether they are truly different constructs, a number of future studies may be proposed. First, qualitative methods may be used to determine what listeners are measuring when they are asked to judge speech acceptability and PLE; for example, what do they think is meant by the terms? What made one speaker more acceptable or require less effort to listen to? What specific qualities of the sample influenced their decisions? Qualitative data obtained from these open-ended questions may help to refine the concept of listener burden and the definition used in future research.

In addition to refining the meaning of PLE and speech acceptability, additional research should also examine the acoustic basis of the speech samples and how these results relate to perceptual outcomes (Maryn et al., 2009). For example, multidimensional scaling of listener responses may contribute to understanding the attributes of PLE, as well as speech acceptability. In addition to acoustic measures, other physiological measures of effort, such as those used in research on the perspective of hearing impaired listeners, could be used to further investigate the objective basis of the perception of PLE (Evitts & Searl, 2006; Rakerd, Franz, & Whearty, 1996; Zekveld et al., 2010).

It may be fruitful to consider whether there is a difference in the relationship between acceptability and PLE given samples of varying intelligibility; perhaps a difference between these dimensions becomes clearer with reduced intelligibility. Samples of lower intelligibility may have reduced acceptability and/or require more PLE, or there may be some critical point beyond which these dimensions are not affected. For example, an examination of the relationship between speech samples featuring a range of intelligibility and their acceptability and PLE scores may reveal the kind of nonlinear relationship found in the study by Beukelman et al. (2011). Verification of intelligibility by objective measures (i.e., transcription) also would strengthen the control of these effects above and beyond those found in the present study.

Finally, the question of experience should also be examined, as it may be that listeners experienced in communicating with alaryngeal speakers, such as speech-language pathologists or spouses, perceive themselves as using less effort to listen to alaryngeal speech than those without such experience. For example, Finizia et al. (1998) found that inexperienced listeners judged TE speech as overall less acceptable than experienced clinicians.

The potential effects of these factors have important clinical and social implications for both individuals with speech and voice disorders and their communication partners. Consideration of listener burden as a treatment outcome could address the problem of individuals whose speech is intelligible, but who may be aware of limitations of communicative success with unfamiliar listeners. An understanding of the factors affecting PLE may guide clinicians in choosing a focus of treatment; for example, a treatment method that simultaneously increases intelligibility and reduces PLE may be more efficient and effective than one that only increases intelligibility. People who frequently encounter individuals with speech or voice disorders (or differences, such as foreign accent) could also receive training in listener strategies that may reduce perceived effort in communication.

5.0 Conclusions

Inexperienced listeners reliably judged sample pairs of TE speech for acceptability of the samples and their own effort in listening to the samples. Although the concept of perceived listener effort correlated strongly with acceptability for inexperienced listeners in this study, there is reason to believe that there are differences in the way listeners interpret the meanings of these dimensions. Future research is suggested to further explore the construct validity of PLE and investigate its relationship to other auditory-perceptual dimensions of speech and voice.

References

ANSI. (1960). ANSI S1.1-1960, Acoustical terminology. American National Standards Institute, New York.

Anderson Gosselin, P., & Gagné, J.-P. (2011). Older adults expend more listening effort than young adults recognizing speech in noise. *Journal of Speech, Language, and Hearing Research, 54*, 944-958. doi:10.1044/1092-4388(2010/10-0069).

Bennett, S., & Weinberg, B. (1973). Acceptability ratings of normal, esophageal, and artificial larynx speech. *Journal of Speech and Hearing Research, 16*, 608-615.

Beukelman, D.R., Childes, J., Carrell, T., Funk, T., Ball, L.J., Pattee, G.L. (2011). Perceived attention allocation of listeners who transcribe the speech of speakers with amyotrophic lateral sclerosis. *Speech Communication, 53*, 801-806. doi:10.1016/j.specom.2010.12.005.

Chan, K. M., & Yiu, E. M. (2002). The effect of anchors and training on the reliability of perceptual voice evaluation. *Journal of Speech, Language, and Hearing Research, 45*, 111-126. doi: 10.1044/1092-4388(2002/009).

Eadie, T. L. (2003). The ICF: a proposed framework for comprehensive rehabilitation of individuals who use alaryngeal speech. *American Journal of Speech-Language Pathology, 12*, 189-197.

Eadie, T. L. (2007). Application of the ICF in communication after total laryngectomy. *Seminars in Speech and Language, 28*, 291-300.

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

Eadie, T. L., & Doyle, P. C. (2004). Auditory-perceptual scaling and quality of life in tracheoesophageal speakers. *Laryngoscope, 114*, 753-759.

Eadie, T. L., & Doyle, P. C. (2005). Quality of life in male tracheoesophageal (TE) speakers. *Journal of Rehabilitation Research and Development, 42*, 115-124.

Eadie, T. L., Doyle, P. C., Hansen, K., & Beaudin, P. G. (2008). Influence of speaker gender on listener judgments of tracheoesophageal speech. *Journal of Voice, 22*, 43-57. doi: S0892-1997(06)00111-1 [pii]10.1016/j.jvoice.2006.08.008.

Eadie, T. L., Nicolici, C., Baylor, C., Almand, K., Waugh, P., & Maronian, N. (2007). Effect of experience on judgments of adductor spasmodic dysphonia. *Annals of Otology, Rhinology and Laryngology, 116*, 695-701.

Evans, D., Healey, E. C., Kawai, N., & Rowland, S. (2008). Middle school students' perceptions of a peer who stutters. *Journal of Fluency Disorders, 33*, 203-219. doi: S0094-730X(08)00034-X [pii]10.1016/j.jfludis.2008.06.002.

Evitts, P. M., & Searl, J. (2006). Reaction times of normal listeners to laryngeal, alaryngeal, and synthetic speech. *Journal of Speech, Language, and Hearing Research, 49*, 1380-1390. doi: 49/6/1380 [pii]10.1044/1092-4388(2006/099).

Fairbanks, G. (1960). *Voice and articulation drillbook* (2d ed.). New York: Harper.

Finizia, C., Lindström, J., & Dotevall, H. (1998). Intelligibility and perceptual ratings after treatment for laryngeal cancer: Laryngectomy versus radiotherapy. *Laryngoscope, 108*, 138-143.

Healey, E. C., Gabel, R. M., Daniels, D. E., & Kawai, N. (2007). The effects of self-disclosure and non self-disclosure of stuttering on listeners' perceptions of a person who stutters. *Journal of Fluency Disorders, 32*, 51-69. doi: S0094-730X(07)00002-2, [pii]10.1016/j.jfludis.2006.12.003.

Iversen-Thoburn, S. K., & Hayden, P. A. (2000). Alaryngeal speech utilization: A survey. *Journal of Medical Speech-Language Pathology, 8*, 85-99.

Klasner, E. R., & Yorkston, K. M. (2005). Speech intelligibility in ALS and HD dysarthria: The everyday listener's perspective. *Journal of Medical Speech-Language Pathology, 13*, 127-139.

Kreiman, J., & Gerratt, B. R. (1996). The perceptual structure of pathologic voice quality. *Journal of the Acoustical Society of America, 100*, 1787-1795.

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

Kreiman, J., Gerratt, B. R., Precoda, K., & Berke, G. S. (1992). Individual differences in voice quality perception. *Journal of Speech and Hearing Research, 35*, 512-520.

Maryn, Y., Dick, C., Vandenbruaene, C., Vauterin, T., & Jacobs, T. (2009). Spectral, cepstral, and multivariate exploration of tracheoesophageal voice quality in continuous speech and sustained vowels. *Laryngoscope, 119*, 2384-2394. doi: 10.1002/lary.20620.

Meltzner, G.S., & Hillman, R.E. (2005). Impact of aberrant acoustic properties on the perception of sound quality in electrolarynx speech. *Journal of Speech, Language and Hearing Research, 48*, 766-779. doi:10.1044/1092-4388(2005/053).

O'Brian, S., Packman, A., Onslow, M., Cream, A., O'Brian, N., & Bastock, K. (2003). Is listener comfort a viable construct in stuttering research? *Journal of Speech, Language, and Hearing Research, 46*, 503-509.

Pindzola, R. H., & Cain, B. H. (1988). Acceptability ratings of tracheoesophageal speech. *Laryngoscope, 98*, 394-397.

Portney, L., & Watkins, M. (2000). *Foundations of clinical research : Applications to practice* (2nd ed.). Upper Saddle River: Prentice Hall Health.

Rakerd, B, Seitz, F., P., & Whearty, M. (1996). Assessing the cognitive demands of speech listening for people with hearing losses. *Ear & Hearing, 17*, 97-106. Retrieved from http://journals.lww.com/ear-hearing/pages/default.aspx

Robbins, J., Fisher, H. B., Blom, E. C., & Singer, M. I. (1984). A comparative acoustic study of normal, esophageal, and tracheoesophageal speech production. *Journal of Speech and Hearing Disorders, 49*, 202-210.

Searl, J. P., & Small, L. H. (2002). Gender and masculinity-femininity ratings of tracheoesophageal speech. *Journal of Communication Disorders, 35*, 407-420.

Singer, M. I., & Blom, E. D. (1980). An endoscopic technique for restoration of voice after laryngectomy. *Annals of Otology, Rhinology and Laryngology, 89*, 529-533.

Trudeau, M. D. (1987). A comparison of the speech acceptability of good and excellent esophageal and tracheoesophageal speakers. *Journal of Communication Disorders, 20*, 41-49.

Trudeau, M. D., & Qi, Y. Y. (1990). Acoustic characteristics of female tracheoesophageal speech. *Journal of Speech and Hearing Disorders, 55*, 244-250.

Whitehill, T., & Wong, C. (2006). Contributing factors to listener effort for dysarthric speech. *Journal of Medical Speech-Language Pathology, 14*, 335–342.

Zekveld, A.A., Kramer, S.E., Festen, J.M., (2010). Pupil response as an indication of effortful listening: The influence of sentence intelligibility. *Ear & Hearing, 31*, 480-490. doi: 10.1097/AUD.0b013e3181d4f251

Appendix

Instructions for listener effort task:

You will be listening to speech samples from adult males. Please rate these samples in terms of LISTENER EFFORT. LISTENER EFFORT is the amount of work needed to listen to a speaker.

Please rate the speech sample pairs for LISTENER EFFORT using the scale provided. You will hear each sample pair only once.

Here is an example of the scale:

Speaker #1 Neutral Speaker #2

To rate the speech sample, please drag the cursor on the scale to indicate which speaker required LESS effort for you to listen to, and by how much. For example, if you perceive Speaker #1's voice to require LESS effort than Speaker #2's, please drag the cursor toward the left end of the scale. If the speakers require an equal amount of effort, please drag the cursor to the middle of the scale ("neutral"). If Speaker #2 requires LESS effort than Speaker #1, please drag the cursor toward to the right. Remember that you may move the cursor anywhere on the scale if you believe it applies. If you believe one speaker demands much LESS effort than the other, move the cursor farther toward that end.

Chapter 3

Utility of perceived listener effort as an outcome measure

Abstract

Perceived listener effort (PLE) is a perceptual dimension used to identify the amount of work necessary to understand speech samples. Relationships among ratings of PLE and speech acceptability for electrolaryngeal speech samples with a range of intelligibility were investigated to provide support for PLE as an outcome measure. Twenty-five inexperienced listeners transcribed and rated samples produced by healthy male speakers using an electrolarynx. Perceived listener effort was highly correlated with intelligibility (r = -0.83) and acceptability (r = -0.88), and was shown to be a multidimensional construct, with the factors of acceptability, intelligibility, and sentence length significantly predicting 86% of the variance in PLE ratings. Acceptability showed both the strongest univariate relationship with PLE as well as a significant unique relationship with PLE when other factors were held constant. The interaction of sentence length and intelligibility on PLE ratings was also significant (p<0.05), with Long (13-15 word) sentences requiring greater effort than equally intelligible Short (5-7 word) sentences. The additional finding that sentence length did not affect acceptability ratings of the same samples suggests that acceptability ratings are made based on a *gestalt* impression of the stimulus, whereas PLE ratings vary as a function of sentence length, among other variables.

1. Introduction

Multidimensional assessment is standard in the evaluation of speech and voice disorders for research, and may include aerodynamic, acoustic and physiological measures. Perceptual quality (i.e., the way disordered speech sounds to listeners) is also of critical importance in determining treatment outcomes. However, instrumental measures may not correlate strongly with perceived voice quality or speech severity. For this reason, numerous perceptual dimensions are used to evaluate disordered speech and voice to ensure a comprehensive approach (Bunton, Kent, Duffy, Rosenbek & Kent, 2007; Kent 1996). For example, perceptual assessment of voice disorders may involve administration of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V; ASHA, 2002; Kempster, Gerratt, Verdolini, Barkmeier-Kramer & Hillman, 2009), which includes ratings of voice qualities such as roughness, breathiness, strain, pitch, loudness and overall severity. For alaryngeal speech, ratings of acceptability and naturalness may also be obtained (Doyle & Eadie, 2005). Finally, for disordered speech, various perceptual ratings of speech may complement measures of intelligibility, which are usually measured as a percentage of total words understood for a given sample.

The personal characteristics of a given listener may influence perceptual ratings, and may also affect the relatively objective measurement of speech intelligibility (Bradlow, Torretta & Pisoni, 1995; Kent, 1996; Kreiman, Gerratt, Kempster, Erman & Burke, 1993). In evaluations of disordered speech, however, little attention has been paid to the role of communication partners and the burden placed on them by disordered speech. The literature of hearing science has a long history of examining how noise and hearing loss can increase the amount of effort needed for listeners to process speech in various contexts. This line of research favors a behavioral approach in which “listening effort” is indexed by performance on reaction time to a second challenging task presented simultaneously with a primary speech recognition task (Gosselin & Gagné, 2010;

McGarrigle, Munro, Dawes, Stewart, Moore, Barry et al., 2014). This is known as a dual-task paradigm. In contrast, investigations of the impact of disordered speech or voice on listener effort have almost exclusively involved self-ratings of attention allocation, mental effort or perceived listener effort (Beukelman, Childes, Carrell, Funk, Ball & Pattee, 2011; Beukelman, Gillespie, Fager & Ullman, 2014; Nagle & Eadie, 2012b; Panico & Healey, 2009; Whitehill & Wong, 2006). Numerous studies using both methods have found that objective measures of "listening effort" do not actually correspond well with subjective measures of perceived mental effort or listener effort (Desjardins & Doherty, 2013; Gosselin & Gagné, 2011; Larsby, Hallgren, Lyxell & Arlinger, 2005; Picou, Ricketts & Hornsby, 2011; Zekveld & Kramer, 2014). Some researchers consider such findings to indicate that objective measures more validly represent the construct of listening effort than subjective measures because they are derived from performance or physiological changes. However, the decision to initiate or maintain a conversation with a speaker with a speech or voice disorder is likely just as strongly influenced by a listener's perception of how hard he or she has to work on listening, as it is by the ultimate success of the interaction. In other words, subjective and objective measures of listening effort may uncover different but equally relevant aspects of listener-speaker interactions.

1.1 Speech intelligibility

The value of perceived listener effort (PLE) as an outcome measure depends on its ability to provide unique information not captured by other measures. For this reason, ratings of listener effort have been compared to measures of speech intelligibility, as well as speech and voice quality. Intelligibility is interpreted as the accuracy with which a listener is able to decode an utterance (Kent, Weismer, Kent & Rosenbek, 1989). It might be expected that as intelligibility increases, the amount of effort required to listen to it would monotonically decrease. After all,

speech acceptability, naturalness and overall severity tend to strongly correlate with intelligibility (Ellis, 1999; Southwood & Weismer, 1993; Nagle, Wright, Sumida, & Eadie, 2012). However, studies investigating the association between listener effort and the intelligibility of dysarthric speech have found that this relationship is not so predictable.

Whitehill and Wong (2006) defined listener effort as the "effort needed to understand the speaker" in their investigation of the perceptual factors contributing to listener effort. Twenty inexperienced listeners transcribed decontextualized sentences and rated listener effort using an undifferentiated 10cm visual analog scale (VAS). The authors grouped 33 Cantonese dysarthric speakers by etiology (i.e., Parkinson's disease, cardiovascular accident, cerebral palsy and "other"), representing hypokinetic, spastic, athetoid, and mixed dysarthrias. Apart from intelligibility scores, no perceptual measures of severity of disorder (i.e., acceptability, naturalness) were provided. As predicted, sentence intelligibility scores and ratings of listener effort were strongly correlated for their samples of Cantonese dysarthric speech (Spearman's R = -0.95). However, three speakers with high intelligibility (greater than 85%) received relatively high mean ratings of listener effort, indicating a distinction between transcribed intelligibility and the perceived effort needed to understand some speech. No particular perceptual features seemed to differentiate the low-effort from the high-effort speakers; in fact, "slurred speech" was among the top three features selected for all. Because results for each dysarthric speech type were not systematically reported, it is difficult to generalize their findings about the relationships among intelligibility, listener effort and specific perceptual features to any one dysarthria type. A similar, but more moderate correlation (rho = -0.57), also was found between perceptual ratings of listener effort and transcriptions of dysarthric speech in children (Coté-Reschny & Hodge, 2010).

Two additional studies by Beukelman and colleagues also have focused on examining relationships between intelligibility and a construct related to listener effort. First, Beukelman et al. (2011) obtained ratings of attention allocation expended by five inexperienced listeners for samples of dysarthric speech. Samples were chosen to represent a range of mild to severe dysarthria associated with amyotrophic lateral sclerosis (ALS). Listener participants transcribed decontextualized sentences as a measure of speech intelligibility. Ratings of attention allocation, which can be reasonably interpreted as synonymous with listener effort, were obtained using a 7-point equal appearing interval (EAI) scale. Consistent with Whitehill and Wong (2006), results indicated a nonlinear relationship between attention allocation and intelligibility. While listener effort decreased with an increase in intelligibility, results were somewhat unpredictable. Somewhat surprisingly, the highest ratings of attention allocation were given to speech samples that were 75% to 80% intelligible. The authors suggested that ratings of attention allocation may have dropped off when speech was less than 75% intelligible because listeners had ceased to expend additional effort when intelligibility was sufficiently low. A strong correlation between intelligibility and attention allocation (r = -0.86) was reported, but only for speakers with intelligibility scores above 75%. Because of the small sample size, no attempt appears to have been made to describe the full range of the relationship. Further, although mean intelligibility scores in this study ranged from 3% to 100%, no mean scores between 40% and 65% intelligibility were reported. In fact, although the authors reported making an effort to represent a majority of intelligibility scores between 60% and 100%, their results appear to indicate that only one speaker's intelligibility score fell between 60% and 75%. By obtaining perceptual ratings from only five listeners and under-representing the range of dysarthric speakers whose mean intelligibility is less than 75%, this study provides useful but limited information about the

relationship between intelligibility and measures of effort.

Beukelman et al. (2014) used similar methods to evaluate perceived attention allocation for speech produced by speakers with traumatic brain injury. Mean intelligibility scores and ratings of perceived attention allocation were derived from 30 listeners, based on sentences produced by 27 speakers. Mean intelligibility scores ranged from 3% - 97% and were more equally distributed than in the previous (2011) study. Although the specific association between attention allocation and intelligibility was not reported, the data as presented in a scatter plot appeared to be nonlinearly related. Below 70% mean intelligibility, mean attention allocation scores ranged from 6-7 on the 7 point EAI scale; however, above 70% mean intelligibility, perceived attention allocation scores were much less predictable. These results support the earlier finding of a nonlinear relationship between intelligibility and perceived attention allocation. They also reveal that, at least for dysarthria associated with TBI, very highly intelligible speech may still require moderate levels of attentional effort to understand.

Despite the limitations of these studies, their overall findings are consistent. The potential discontinuity between speech intelligibility scores and ratings of listener effort suggests that PLE addresses aspects of speech signal processing beyond simple decoding of the signal.

1.2 Perceptual dimensions of disordered speech

Intelligibility is commonly measured for both speech and voice disorders, but the perceptual parameters related to a particular disorder type may provide more meaningful descriptions of an individual's condition (Bunton et al., 2007; Eadie & Doyle, 2005; Law, Ma & Yiu, 2009). Ratings of naturalness, acceptability or overall severity are routinely used to more fully describe disordered speech and voice (e.g., Eadie & Doyle, 2005; Southwood & Weismer, 1993). When completely intelligible speech contains obvious distortions or differences that may

be distracting to a listener, these perceptual parameters provide a means of documenting changes in speech or voice quality. For example, Sussman and Tjaden (2012) determined that perceptual ratings of overall severity were more sensitive than measures of intelligibility to differences in the speech of patients with multiple sclerosis, Parkinson's disease and healthy controls. Intelligibility was measured in terms of both percent single words correct (based on a four-alternative forced choice task) and words in sentences correct from the *Sentence Intelligibility Test* (SIT; Yorkston, Beukelman & Tice, 1996). Overall severity was measured using a 150 mm VAS, marked at endpoints only (*no impairment* to *severe impairment*). The authors intended to establish whether intelligibility scores fully captured listener impressions of "speech adequacy." Despite their use of different listening paradigms and speech stimuli for each condition, the results clearly demonstrated that global perceptual ratings like overall severity contributed information beyond relatively objective measures of intelligibility.

The utility of PLE as a clinical descriptor depends on its ability to provide independent information about a speech signal, beyond measures already used to describe the signal's auditory-perceptual characteristics. Speech disorders usually affect production accuracy to some degree, making intelligibility scores a satisfactory standard against which to judge the utility of listener effort ratings. Considered in terms of the source-filter model of speech production, it makes sense that a damaged set of articulatory "filters" would affect intelligibility, whether or not the voice "source" was intact. Disorders of voice, however, are generally considered to affect the speech "source." Unless a disorder of voice is severe, it will tend to have minimal effects on intelligibility; rather, the distracting effects of a disordered voice tend to affect the perceptual

qualities of the voice.[1] Voice evaluations, therefore, tend to focus on the perceptual source characteristics of the signal rather than its intelligibility. Perceptual judgments of voice quality remain the standard against which information about acoustics, aerodynamics, physiological function and quality-of-life are considered (Kreiman et al., 1993).

The particular perceptual dimensions used to measure clinical outcomes differ for laryngeal and alaryngeal voices. For alaryngeal speech, acceptability is one of the most common perceptual dimensions measured (e.g., Eadie & Doyle, 2005; Pindzola & Cain, 1988); for laryngeal-based voice disorders, measurement of "overall severity of disorder" is more common (e.g., Eadie & Kapsner-Smith, 2011; Stewart, Allen, Tureen, Diamond, Blitzer & Brin, 1997). Whereas "overall severity" is likely to capture differences among voices produced with a laryngeal source, the obvious differentness of voices produced with a pharyngo-esophageal or mechanical source makes "acceptability" a more reasonable dimension for alaryngeal voice outcomes. To date, only two studies have compared perceptual dimensions of speech quality to PLE.

Twenty listeners rated PLE and acceptability of tracheoesophageal (TE) speech (Nagle and Eadie, 2012a). Because an identical elicitation stimulus was used for all samples (second sentence of the Rainbow Passage [Fairbanks, 1960]), listeners knew the target for each sample. Samples represented the full range of acceptability, as judged by expert listeners. Listener reliability was maximized by using a paired samples paradigm, in which each sample was judged relative to every other sample. In this way, each judgment was made relative to an "external" standard. Each sample pair was rated using a 100mm VAS on which listeners could indicate the

[1] The notable exception to this presumed source vs. filter dichotomy is alaryngeal speech, which is produced through an oropharyngeal filter that has been modified through surgery, chemotherapy, radiation or a combination of these treatments.

degree of difference between the samples in each pair. Mean perceptual ratings of listener effort were strongly correlated with judgments of acceptability (r = .99).

The previous study was partially replicated using samples of adductor spasmodic dysphonia speech (ADSD; Nagle & Eadie, 2012b). Rather than rating acceptability, listeners rated overall severity, because ADSD is a laryngeal-based neurological voice disorder. Listeners rated both overall severity and listener effort for paired samples of ADSD speech. As in the previous study, the perceptual dimension of speech quality was highly correlated with listener effort for ADSD speech (r = .98).

1.3 Research questions

1.3.1 Interrelationships among outcome measures

Thus far, two patterns have been identified in the relationships between PLE and other perceptual outcome measures of speech and voice. On one hand, PLE and speech intelligibility appear to be nonlinearly related, with a reported peak of PLE at relatively high but not perfect levels of intelligibility (i.e., 70%-85%; Beukelman et al., 2011). On the other hand, PLE and the perceptual dimensions of speech acceptability (for TE speech) and overall severity (for ADSD speech) appear to be very strongly correlated, at least for group data obtained using paired samples (Nagle & Eadie, 2012a; 2012b).

Two possibly complimentary propositions suggest themselves. First, it could be that PLE is associated more strongly with other perceptual dimensions, such as overall severity or acceptability, than with intelligibility. In the absence of varied intelligibility, Nagle and Eadie (2012a; 2012b) found strongly linear correlations between these dimensions of voice quality and PLE. Findings of a nonlinear relationship between PLE and intelligibility for dysarthric speech are consistent with this possibility (Beukelman et al., 2011; 2014). In short, it may be that PLE is

strongly related to intelligibility, but its relationship with other perceptual dimensions may be stronger. If so, introducing PLE as a clinical consideration for speech and voice disorders may not be useful.

A second possibility arising from the results of the perceptual studies is that stimulus type may affect the relationships among these outcome measures (Nagle & Eadie, 2012a; 2012b). Relatively speaking, more categorical changes between dimensions were noted for TE speech (PLE vs. acceptability) than for ADSD (PLE vs. overall severity). Far fewer ratings were equivocal for the TE speakers; mean ratings were well-distributed for both acceptability and listener effort. For ADSD, on the other hand, mean ratings were distributed within the low range for both overall severity and listener effort. This disparity could be related to a difference in sample size (n=14 TE samples vs. n=10 ADSD samples), or it could be that voice source (laryngeal vs. alaryngeal) is an important factor in finding differences between these dimensions. For example, the shapes of the functions describing the relationship between intelligibility and listener effort could be quite different for electrolaryngeal speech when compared to the nonlinear function described by Beukelman et al. (2011; 2014). The relationship between intelligibility and listener effort seems likely to vary, depending on numerous variables including speaker and listener population, range of disorder severity and stimulus complexity, such as sentence length.

The research described this far has reported correlations between PLE ratings and measures of either speech intelligibility or one other common perceptual dimension. As noted above, peak listener effort has been reported to be within 70-85% intelligibility, for the dysarthria speech associated with ALS, but the range between 40-60% intelligibility has not yet been well sampled (Beukelman et al., 2011; 2014). The current study was designed to examine

samples representing the entire range of intelligibility scores, but the range between 40-95% intelligibility was of particular interest. The interrelationships among PLE, intelligibility and other perceptual outcome measures have not yet been investigated for disordered speech. More data are needed to establish the relationships among auditory-perceptual outcome measures and intelligibility if findings are to be generalized to other populations and levels of severity. Given findings on the relationship between speech samples produced with an identical elicitation stimulus, it is important to use speech samples with a range of intelligibility (and acceptability) to clarify the relationship between PLE and intelligibility. As a result, the primary question of this study was:

1. How do ratings of perceived listener effort compare to those for speech acceptability for speech samples with a wide range of intelligibility?

1.3.2 Effects of varying signal-dependent factors

Another way to examine the independent contribution of PLE as an outcome measure is to systematically vary the characteristics of the stimuli to be rated. For example, under certain circumstances, speech samples with similar intelligibility scores and acceptability ratings might have different effects on PLE ratings. Likewise, speech samples of varying length or complexity may elicit different ratings of PLE when intelligibility and acceptability are controlled.

Current models of working memory indicate limited cognitive "space" available for simultaneous processing of complex stimuli. For example, longer or more complex auditory stimuli require more space in the working memory buffer, leading to reduced available processing capacity (Baddeley, 2000). Numerous studies from the hearing aid literature have supported the idea that decreased signal-to-noise ratio (SNR) leads to reduced available working memory capacity, as demonstrated by increased processing time for dual vs. single tasks (Fraser,

Gagne, Alepins & Dubois, 2010; Gosselin & Gagne, 2011). Effects of stimulus complexity, however, have not yet been examined systematically for ratings of listener effort; for example, effects of utterance length, linguistic complexity, and content are unknown. Even in the more substantial hearing literature on perceptual ratings of listener effort, stimuli have generally been limited to a single level of stimulus complexity per experiment. For example, Rudner, Lunner, Behrens, Thorén and Rönnberg (2012) presented only 5-word sentences following a strict noun-verb-number-adjective-noun pattern to examine the effects of noise on working memory. Ton, McCoy and Wingfield (2009) investigated the effects of age and hearing acuity on recall using semantically related versus unrelated words. These stimuli are relatively simple and similar in length because they were controlled to be so, but they have little real-world application. Longer, less grammatically controlled stimuli would approximate real-world listening conditions more closely. It is therefore hypothesized that increasing the length of an utterance should also cause a perception of increased effort required to process it, without necessarily decreasing its intelligibility or acceptability. Evaluating differences in perceptual ratings related to stimulus length is a first step in testing the effects of signal-dependent factors on PLE. Consequently, the second and third questions answered in this study were:

2. What is the effect of sentence length on ratings of perceived listener effort across a range of acceptability and intelligibility?

3. Does sentence length affect ratings of perceived listener effort and acceptability equally for equally intelligible speech?

2. Methods

This study investigated the relationships among PLE ratings, acceptability and intelligibility using samples of electrolaryngeal (EL) speech obtained from healthy male

speakers. EL speech is particularly suitable for the initial investigation of these relationships because it is easy to produce a set of speech stimuli with a range of intelligibility. Findings from this study have potentially immediate effects on current users of EL speech. First, the intelligibility of EL speech varies widely (Kalb & Carpenter, 1981; Weiss & Basil, 1985; Meltzner & Hillman, 2005); even without added noise masking, ceiling effects on intelligibility are unlikely. Second, both source and filter characteristics of EL speech can be controlled naturally, without artificial manipulation of the samples. The single EL device voice "source" is typically limited to a single fundamental frequency. Similarly, laryngopharyngeal morphology among individuals who have undergone total laryngectomies may be unpredictably varied from the effects of treatment or surgery, causing unknown interspeaker differences in "filter" effects on speech; using trained healthy speakers of similar sex and age can limit these differences. Finally, EL speech remains a commonly used mode of communication for alaryngeal speakers, both as a primary and backup method (Doyle, 2005; Hillman, Walsh, Wolf, Fisher & Hong, 1998). The unnatural or robotic sound of EL speech is largely due to radiated noise from the device and occurs to some degree regardless of the health of the vocal tract. Thus, although the current study used samples produced by healthy speakers, results have implications for individuals who use EL speech as their primary method of communication.

2.1 Participants

All participants (speakers and listeners) were native speakers of American English, with no history of a speech, language or hearing disorder. Apart from age requirements, there were no exclusionary criteria for this study. All participants were paid for their participation in this study. All procedures were approved by the Institutional Review Board at the University of Washington.

2.1.1 Speakers

Ten healthy males between ages 50 and 65 were recruited to provide a corpus of speech samples. Their laryngeal speech quality was judged as normal by the first author, an experienced speech-language pathologist, immediately preceding recording.

2.1.2 Listeners

Two groups of inexperienced listeners between ages 18 and 35 were used in this experiment. Although older listeners may be more likely to interact with alaryngeal speakers, there is evidence that older listeners report subjective listener effort scores that align with objective measures of cognitive effort even less than those from younger listeners (Desjardins & Doherty, 2013; Gosselin & Gagné, 2011; Larsby et al., 2005). Using younger listeners is appropriate for a study of inexperienced listeners, because they are less likely to have encountered alaryngeal speakers in their daily lives and because their data are not complicated by age-related hearing issues. All listeners passed a hearing screening at the octave frequencies between 250 Hz and 4kHz at 25 dB SPL. The task of the first group of listeners (n=5) was to transcribe all speech samples in the corpus to ensure that a range of intelligibility was represented in the main experiment. The second group of listeners (10 male, 15 female) participated in the main experimental tasks.

2.2 Procedures

2.2.1 Stimulus preparation

Ten speakers were trained to use a single EL device (Solatone, Griffin Laboratories), calibrated at 75 Hz. This relatively low pitch setting is typical for male alaryngeal speakers and has been shown to enhance the intelligibility of EL speech, relative to higher frequencies (Globleck, Stajner-Katusic, Musura, Horga & Liker, 2004; Nagle, Eadie, Wright & Sumida,

2012). Speakers were instructed in the use of the EL device, specifically in the proper placement of the EL diaphragm against the neck. After several minutes of practice, they were given a randomly assigned list of sentences from the *Assessment of Intelligibility for Dysarthric Speech* (AIDS; Yorkston & Beukelman, 1984), which is the hard-copy version of the computerized SIT. For recording, they were specifically told to produce all of the words in each sentence; for example, they were told not to contract verbs unless the stimulus sentence was written with contractions. Each speaker was recorded reading the entire assigned list of sentences aloud, with opportunities to re-record samples as requested. The first author monitored recording and ensured that stimuli were produced as written.

To determine whether sentence length affects the relationship between intelligibility and listener effort, speech stimuli included both short and long sentences. All stimulus sentences used to create the sample corpus had relatively low semantic predictability. Specifically, stimulus sentences were less than 50% predictable, based on methods used by Beverly, Cannito, Chorna and Bene (2010; M. Cannito, personal communication, March 2013). Each speaker produced 20 sentences in each sentence-length condition (Short = 5-7 words; Long =13-15 words). Thus, 40 sentences were produced by each speaker (20 x 2 sentence lengths).

Speakers were recorded individually in a quiet room. A free-standing condenser microphone was positioned twelve inches from the mouth of the speaker, and connected to an amplifier (Apogee Digital Trak 2). Samples were recorded at a sampling rate of 44 kHz with 16-bit quantization, and obtained on a desktop computer using a specialized sound card and acoustic software (Sony Soundforge 10.0). Fundamental frequency of 75 Hz ±3Hz was verified for each stimulus after recording using acoustic software (Praat Version 5.3.56, Boersma & Weenink, 2014). Each stimulus was edited and RMS normalized for peak intensity at -24 dB: this

maximized the overall volume of the signal and reduced signal-to-noise ratio while avoiding clipping.

A small set of inexperienced listeners (n=5) transcribed all sentences in the sample corpus, establishing a mean intelligibility score for each sentence (10 speakers x 40 sentences = 400 samples in corpus). Obtaining a large number of stimuli ensured that the purported peak range of high listener effort would be well-represented in the stimuli used for the rating experiment. Intelligibility was established using the protocol described by the AIDS, which allows one repetition of each stimulus, for a total of two plays per listener. Listeners orthographically transcribed the sentences on a desktop computer. They were instructed to type exactly what they heard and to make their best guess if uncertain (Yorkston & Beukelman, 1984). Intelligibility was calculated using a total word phonemic match scoring model, as described in Hustad and Cahill (2003). Transcribed words were counted as correct if all phonemes included in the spelling matched the target word, including homonyms and misspelled words. Listeners were told to write exactly what they heard, being especially careful with plural and contracted forms. Intelligibility was scored as the number of words identified correctly divided by the number of words possible for each trial.

Sentences were chosen from the sample corpus to represent the range of intelligibility of particular interest for this study. The experimental corpus consisted of 244 samples pseudo-randomly assigned to sets of 45-50 sentence samples, each of which had approximately equal numbers of Short and Long sentences. Each stimulus was thus transcribed and rated by five new listeners on each dimension.

2.2.2 Experimental listening procedure

The second listener group comprised five sets of five inexperienced listeners (n=25). This group provided intelligibility scores and ratings of listener effort and acceptability for the samples used in the experiment. Each listener heard one set of 45-50 sentences presented over headphones (Samson Stereo Headphones, RH600), adjusting the volume as desired. A custom-made software program (Matlab R2013a; The Mathworks, Inc., Natick, MA) was used to randomize the presentation order of speech samples. This program allowed the listener to transcribe the speech sample (intelligibility), and recorded the listener responses on a rating scale (listener effort, acceptability). Immediately after transcribing each sample, listeners rated either listener effort or speech acceptability for that sample. PLE was defined as the amount of effort needed to understand a speech sample (Whitehill & Wong, 2006). When rating speech acceptability, listeners were asked to "Give careful consideration to the attributes of pitch, rate, understandability, and voice quality. In other words, is the voice acceptable to listen to as a listener?"(Bennett & Weinberg, 1973, p. 610). Use of these definitions permits comparison of results with previous literature on both listener effort (Nagle & Eadie, 2012a; Nagle & Eadie, 2012b) and acceptability of EL speech (Bennett & Weinberg, 1973; Most, Tobin & Mimran, 2000). Based on how everyday listeners had described rating acceptability in a pilot study, the following guidance was also given to listeners: "Acceptability can be thought of as naturalness, pleasantness and the degree to which the voice is not distracting."

Perceived listener effort and speech acceptability have been shown to be inversely correlated, at least for tracheoesophageal speech (Nagle & Eadie, 2012b). Use of a perceptual scale can be complicated when the endpoints signify different extremes of preference; either one end of the scale can indicate that a sample both required more effort and was more acceptable,

which is counterintuitive, or the extremes can be at opposite ends of the scale, which can be confusing. To simplify instructions to listeners, ratings were elicited in terms of amount or degree of each perceptual construct. Ratings were obtained using a 100mm vertical VAS marked at the endpoints only (e.g., 0 = very little effort/acceptability, 100 = extreme effort/acceptability; Appendix). The bottom of the scale (0mm) represented "less" listener effort and acceptability; the top of the scale suggested "more" effort and acceptability. Using a vertical scale alleviated the potential effects of handedness, which has been shown to affect raters' use of horizontal scales to a greater degree than vertical scales (Chapanis & Gropper, 1968).

Listener raters attended two experimental sessions. At the beginning of each experiment, they were familiarized with the tasks using three speech samples not included in the experiments. This was meant to ensure that listeners not only understood the task, but that they demonstrated a degree of consistency in performing it before experimental data were taken. In the first session, which took about an hour, all listeners transcribed and rated one of the five sets of samples; half of the listeners rated listener effort and the other half rated acceptability. In the second session between one and three weeks later, they judged the remaining dimension for their assigned sample set. This session lasted 20-30 minutes. The delay between the first and second session was meant to control learning effects that might occur if both dimensions were judged on the same day.

2.2.3 Data analysis

To answer the first experimental question, mean intelligibility scores, ratings of acceptability and PLE were first calculated for all stimuli meeting reliability criteria (see below). The associations among intelligibility, ratings of PLE and acceptability and sentence length were determined by calculating Pearson's correlation coefficients. These associations were then

displayed graphically, with regression functions performed to find the lines of best fit for the data in each relationship.

To answer the second and third experimental questions, simultaneous multiple linear regression was conducted to find the effects of intelligibility, acceptability and sentence length (predictor variables) on ratings of PLE (criterion variable). Metrical scores were standardized and entered into a model with an effect-coded categorical length variable and an interaction term combining standardized intelligibility scores and length group. To compare the effects of sentence length on ratings of PLE with effects on acceptability (for experimental question 3), a second simultaneous multiple linear regression analysis was run with acceptability as the criterion variable; predictor variables were standardized intelligibility scores, length group and the interaction term combining standardized intelligibility scores and length group.

All analyses were performed using IBM SPSS Statistics, version 19 (IBM Corporation, 2012).

2.2.3.1 Reliability

Intra-rater agreement was established by comparing ratings for 5 samples (10%) within each set that were repeated for the second dimension rated. In this way, intra-rater agreement was established for 13 listeners based on acceptability ratings and 12 listeners based on listener effort ratings. Two levels of agreement were calculated based on criteria established by Kreiman and Gerratt (1998) for exact agreement and agreement within 1 point on a 7-point equal-appearing-interval scale, which is commonly used in clinic and research. Exact agreement indicated two scores differing by 7.14mm or less, with 14% probability of agreement due to chance. The criterion for agreement within one scale rating was a difference of no more than 21.5mm, with chance agreement at 39%.

Data from five listeners were excluded from analysis for this experiment because their intra-rater agreement was less than chance for either of the levels of agreement. One other listener was excluded because review of her ratings indicated that she may have misunderstood the directions or was not paying attention. Thus, 19 listeners (at least three per sample) contributed data for further data analysis. Overall mean intra-rater exact agreement (within 7.14mm) was 33.77% (range 20-80%), which was well above the chance level of 14%. Agreement within 1 scale point (21.5mm) was 66.44% (range 40-100%), also well above the chance level of 39%. Nine listeners rated listener effort for repeated samples, with mean exact agreement at 40%, and mean agreement within one scale point at 73%. Ten listeners rated acceptability for repeated samples, with mean exact agreement at 38%, and mean agreement within one scale point at 64%. All of these levels of agreement were acceptable, as they were well above the level of chance, and meet criteria for acceptability in perceptual research, with most values around 70% agreement within one scale point on a 7-pt scale (Kreiman et al., 1993).

Inter-rater reliability for both perceptual dimensions was established using intraclass correlation coefficients (ICCs) using a 2-way mixed model in which raters are considered a fixed factor (McGraw & Wong, 1996). When this mixed model is used, inferences should not be made beyond the specific set of raters used within the set. The average measures ICC(A,*k*) provides the mean reliability for all ratings within each set, based on the relation between a single rating and the mean of all ratings for that sample. Criteria for use in further data analysis was $r = .60$ for both dimensions.

Table 3.1. Interrater reliability (ICCs) by sample set.

Sample Set	PLE	ACC
1	0.40	0.72
2	0.73	0.81
3	0.79	0.82
4	0.77	0.64
5	0.56	0.43
Unadjusted Mean (SD)	**0.65 (0.17)**	**0.68 (0.16)**

As is shown in Table 3.1, listeners in sets 1 and 5 did not meet the criteria for interrater reliability and stimuli in these blocks were removed from further analysis. The adjusted overall average measures ICC was .760 for PLE and .755 for acceptability, which indicate good reliability (Portney & Watkins, 2000). With sets 1 and 5 removed, data from 11 listeners (6 male; average age = 22.17 yrs) and 149 samples remained for analysis.

Table 3.2. Ratings per sample for sample sets meeting reliability criteria.

Sample Set	Number of Samples	Ratings per Sample
2	49	3
3	50	4
4	50	4
	N = 149	**11 raters**

3. Results

3.1 Associations among perceived listener effort, acceptability & intelligibility

Pearson's correlation coefficients were calculated for the associations among listener effort, acceptability, intelligibility and stimulus length group (Table 3.3). Perceived listener effort was strongly negatively correlated with acceptability (r = -.88), with 78% of variance shared, and intelligibility (r = -.83), with 70% shared variance. Acceptability and intelligibility scores also were highly positively correlated (r = .81), with 66% shared variance. Notably,

sentence length group was not significantly correlated with either intelligibility scores (r =-.07, p = .212) or acceptability scores (r = .08, p = .183), but it was significantly correlated with PLE (r = .25, p = .001). Relationships between PLE and acceptability ratings as a function of intelligibility scores are displayed in Figures 3.1 and 3.2.

Table 3.3.

Correlation Table

	M	*SD*	*n*	PLE	ACC	INT	Length
1. PLE	6.27	2.40	149	--			
2. ACC	3.66	2.20	149	**-.88** ***	--		
3. INT	0.56	0.31	149	**-.83** ***	**.81** ***	--	
4. Length	0.50	0.50	149	**.25** **	-.08	-.07	--

Note. PLE = perceived listener effort; ACC = acceptability; INT = intelligibility; Length =sentence length group

* p < .05, ** p < .01, *** p < .001.

Figure 3.1. Perceived listener effort as a function of intelligibility scores.
Note: PLE = perceived listener effort

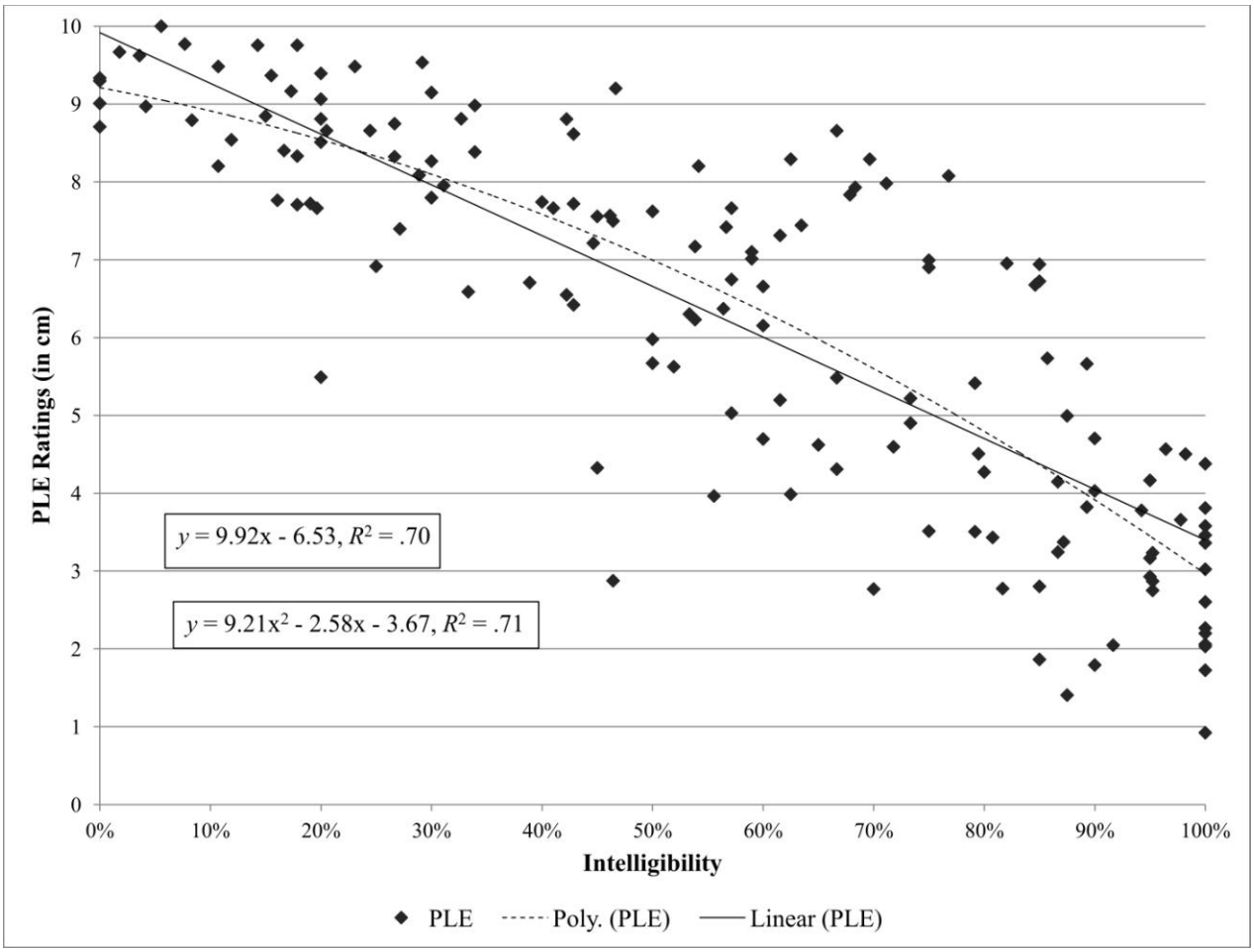
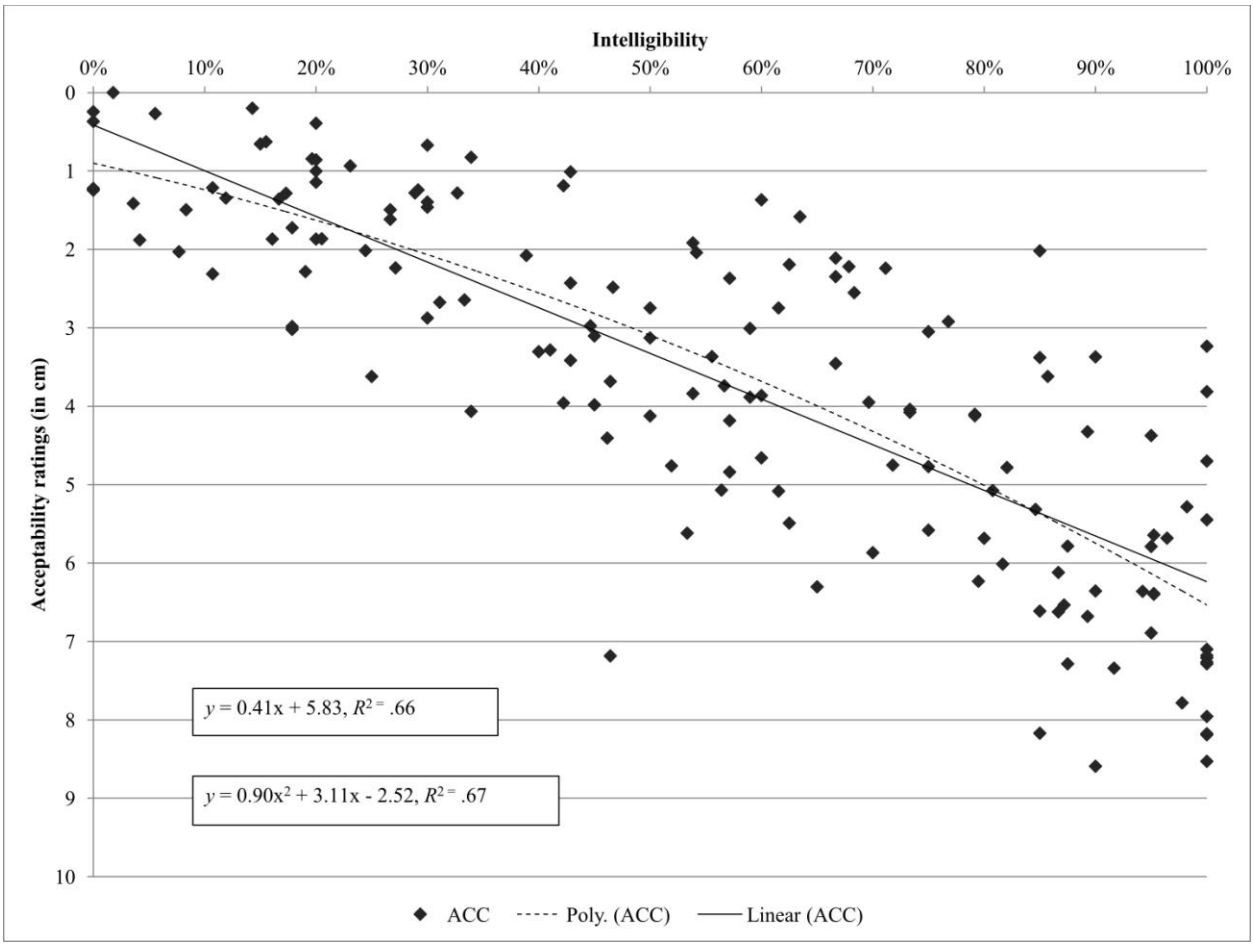


Figure 3.2. Acceptability as a function of intelligibility scores.
Note: ACC = acceptability



3.2 Effect of sentence length on perceived listener effort

Assumptions of normality, linearity and homoscedasticity were met for standard linear regression. The predictors, intelligibility, acceptability ratings and sentence length, accounted for a significant amount of variance in ratings of PLE, $R^2 = .86$, $F(4,144) = 223.84$, $p < .0001$, $R^2_{\text{adjusted}} = .86$ (Table 3.4). Model results are as follow:

$$\bar{Y}_{\text{PLE}} = 6.14 - .79*\text{StdACC} - 1.42*\text{StdINT} + 0.45*\text{Length} + 0.19*\text{StdINT}*\text{Length}$$

The model estimate of the intercept showed that, holding all other variables at average, the mean estimate PLE rating was 6.29 cm ($SE = 0.07$).

Table 3.4.

Multiple Linear Regression (PLE)

	Standard Regression						
	R^2_{total}	R^2_{Adj}	F_{total}	*b*	*(SE)*	*t*	sr^2
PLE	*.86*	*.86*	*223.84(4,144)****				
Intercept				6.14	(0.07)	84.58 ***	
Std ACC				-0.79	(0.13)	-11.14 ***	0.12
Std INT				-1.42	(0.13)	-6.14 ***	0.04
Length				0.45	(0.07)	5.99 ***	0.03
Length x StdINT				0.19	(0.08)	2.49 **	

Note. Std = standardized PLE = perceived listener effort; ACC = acceptability; INT = intelligibility; Length = sentence length group

* $p < .05$, ** $p < .01$, *** $p < .001$.

Ratings of acceptability had a significantly negative effect on ratings of PLE, contributing 12% uniquely to the model ($b = 0.45$, $SE = 0.07$), $t(144) = 5.99$, $p < .001$, $sr^2 = 0.12$. Specifically, there was an estimated mean decrease of .79 cm in PLE rating for each cm increase in acceptability rating, holding other predictors at average.

Intelligibility had a significant negative effect on PLE ratings as well ($b = -1.42$, $SE = 0.13$), $t(144) = -6.14$, $p < .001$, $sr^2 = 0.03$, but this effect was not unique. Specifically, there was an estimated mean decrease of .14 cm in PLE rating for each percent increase in intelligibility, holding other predictors at average.

Sentence length had a significant but not unique positive effect on PLE ratings, ($b = 0.45$, $SE = 0.07$), $t(144) = 5.99$, $p < .001$, $sr^2 = 0.03$. Specifically, there was an estimated mean increase of .45 cm in PLE rating for Long sentences compared to Short sentences, given average intelligibility and acceptability scores. The interaction between intelligibility and length group was also significant for PLE ($p = .01$). Specifically, PLE scores in the Long sentence group were predicted to average 6.14 + 0.19 = 6.33 cm, holding other predictors constant; whereas mean

predicted PLE scores were 6.14 – 0.19 = 5.95 cm for Short sentences. In other words, as intelligibility increased, ratings of PLE decreased, but this effect was stronger for samples in the Short group (Figure 3.3).

Figure 3.3. Association between perceived listener effort and intelligibility for Short and Long sentence levels. *Note:* PLE = perceived listener effort

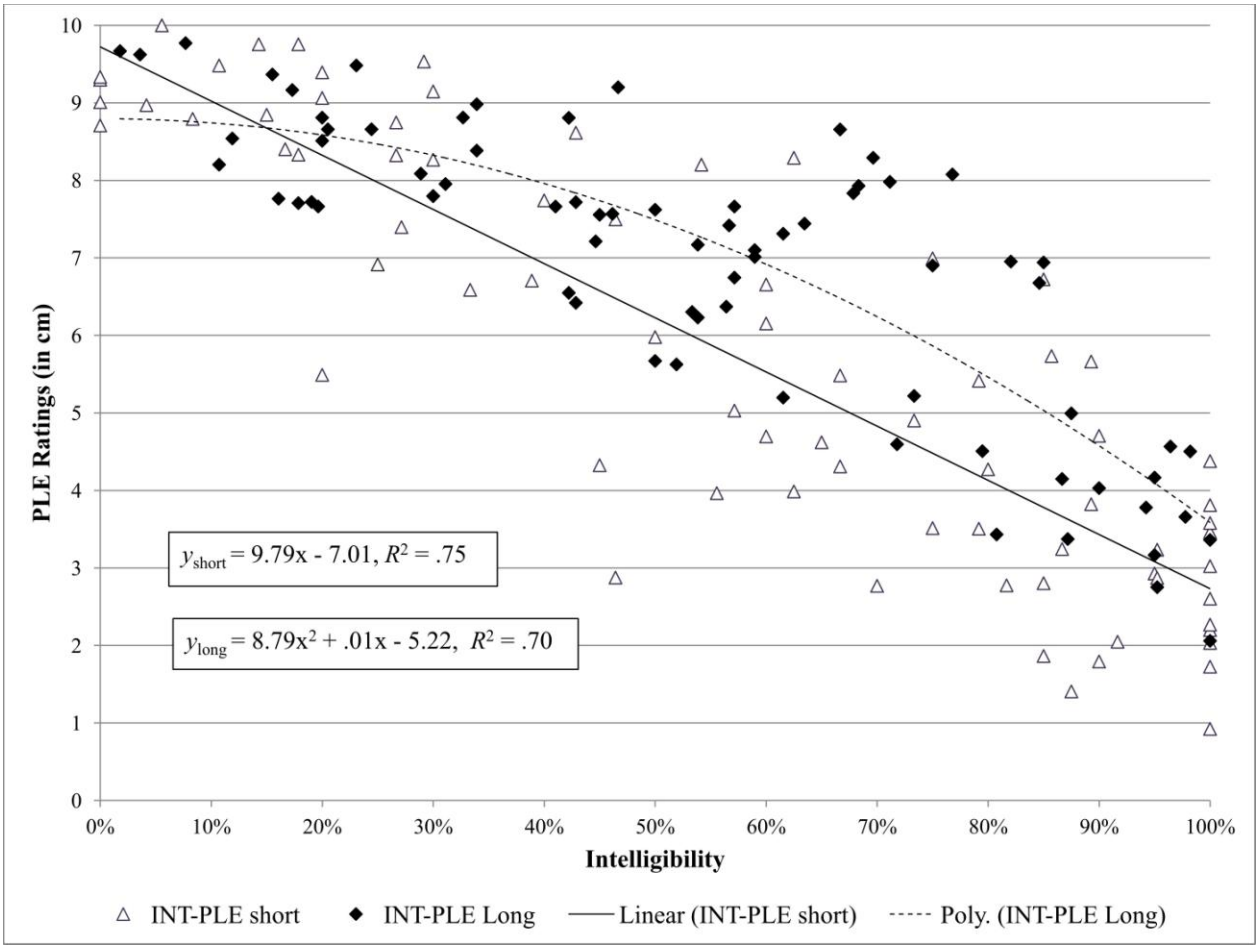
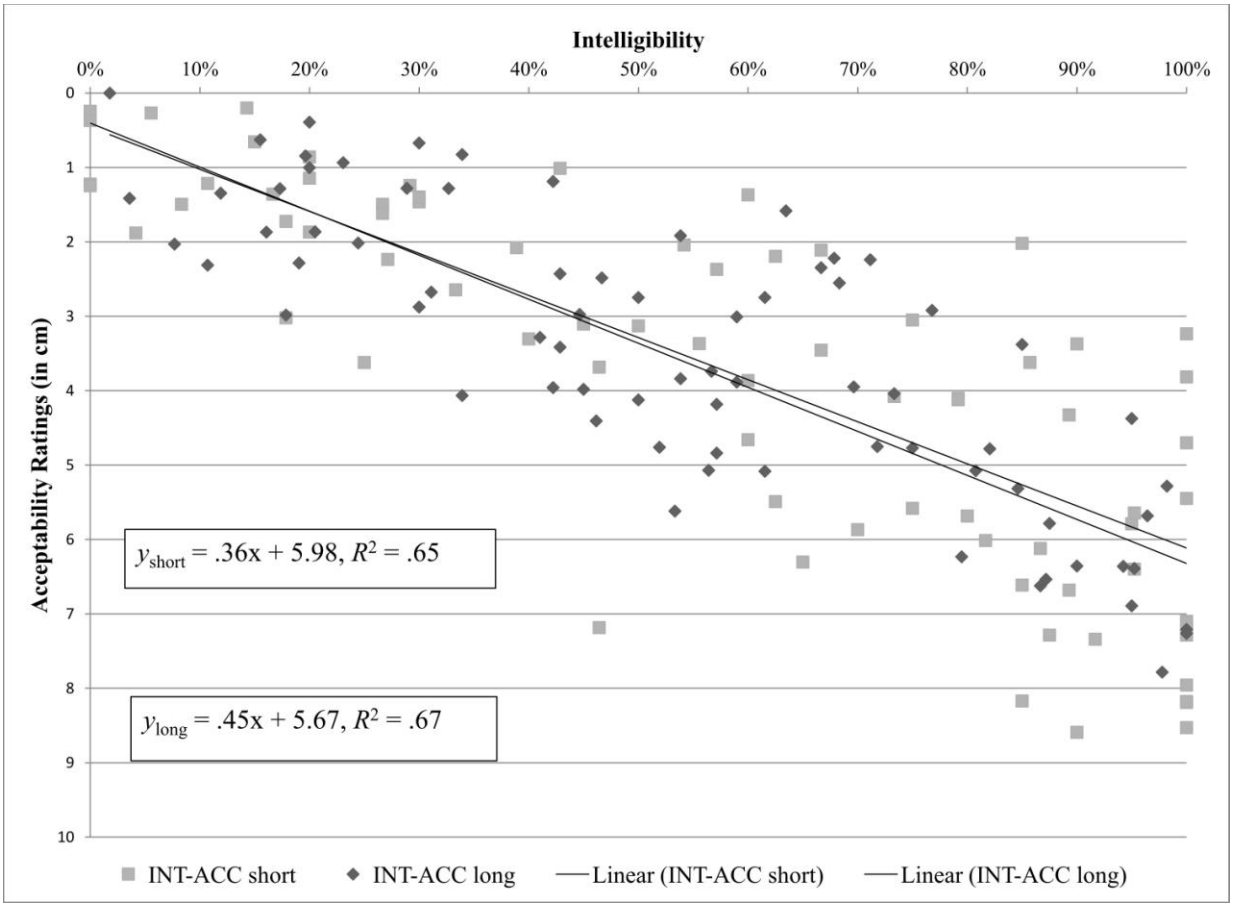


Figure 3.3 displays the association between mean ratings of PLE and intelligibility, with samples grouped by sentence length. The line of best fit for the Short sentence level was linear, R^2 = 0.75. For the Long sentence level, however, the line of best fit was a nonlinear function. The second order polynomial (quadratic) function significantly fit the data ($y = 8.791x^2 + .012x - 5.217$, $r^2 = .71$, F (2,71) = 84.41, $p < .001$, accounting for variance in the data beyond that accounted for by the simple linear model, ($y = 9.899x - 5.6$, $r^2 = .66$, $F(1,72) = 142.43$, $p<.001$) .

3.3 Effect of sentence length on acceptability

The interaction between intelligibility and length group was not statistically significant for ratings of acceptability (p = .723). The strong similarity of the relationship between acceptability and intelligibility for Short and Long sentences is displayed in Figure 3.4. The line of best fit for both levels of sentence length was linear, with R^2 = 0.65 for Short sentences and R^2 = 0.67 for Long sentences.

Figure 3.4. Association between acceptability and intelligibility for Short and Long sentence levels. *Note:* ACC = acceptability



4. Discussion

Ratings of perceived listener effort (PLE) and acceptability of electrolaryngeal speech were obtained for samples with a range of intelligibility. Overall findings were consistent with previous investigations of the association between PLE and other auditory-perceptual dimensions

typically measured for disordered speech (Nagle & Eadie, 2012a; 2012b). That is, there were strong relationships found among the dimensions of intelligibility, acceptability, and PLE. Perceived listener effort was shown to be a multidimensional construct, with the factors of acceptability, intelligibility, and sentence length significantly predicting 86% of the variance in PLE ratings. Acceptability showed both the strongest univariate relationship with PLE as well as a significant unique relationship with PLE when other factors were held constant. The interaction of sentence length and intelligibility on PLE ratings was also significant ($p<.05$), with Long (13-15 word) sentences requiring greater effort than equally intelligible Short (5-7 word) sentences. The additional finding that sentence length did not affect acceptability ratings of the same samples suggests that acceptability ratings are made based on a *gestalt* impression of the stimulus, whereas PLE ratings vary as a function of sentence length, among other variables.

4.1 Interrelationships among measures

The current study is the first to simultaneously compare PLE with both the relatively objective measurement of transcribed intelligibility and an auditory-perceptual dimension of speech quality. Previous studies have found strong correlations between PLE and intelligibility (Beukelman et al., 2011; 2014), acceptability (Nagle & Eadie, 2012b) and overall severity (Nagle & Eadie, 2012a). To show the potential utility of developing a scale of PLE as an outcome measure for disordered speech, listener data were obtained for a large number of individual speech samples with a full range of intelligibility. Two propositions were directly addressed in the current study: that associations of PLE with other perceptual dimensions might be stronger than with intelligibility, and that speech type or sample size might affect those associations.

4.1.1 Auditory-perceptual ratings and intelligibility

The results of this study showed a strong relationship between PLE and speech acceptability. Ratings of acceptability accounted uniquely for 12% of the variance in PLE scores, holding all other factors constant. Investigations of the associations between PLE other auditory-perceptual dimensions of speech or voice quality are few, but results have been consistent with the findings of the current study. Nagle and Eadie (2012b) reported a strong negative correlation between PLE and acceptability for tracheoesophageal speech; they also found a strong positive correlation between PLE and overall severity of speech for ADSD speech (2012b). Brons, Houben and Dreschler (2013) compared perceptual scores across four different hearing aids, obtaining ratings of PLE, "noise annoyance, speech naturalness" and "overall preference" for each hearing aid at 4 levels of intelligibility, but they calculated the relationships between only PLE and intelligibility and overall preference. They reported that noise annoyance and speech naturalness contributed equally to overall preference for hearing aid, but that overall preference was not correlated with either intelligibility or PLE ratings.

Figures 3.1 and 3.2 show the relationships between the two perceptual dimensions used in this study and mean transcribed intelligibility scores. Correlations between both perceptual measures and intelligibility were very strong. There was a minor improvement in R^2 for the line of best fit using a quadratic versus a linear regression function for both dimensions; however, the parsimonious interpretation of the data suggests that, overall, both are associated linearly with intelligibility. On the other hand, visual inspection of the intelligibility data in these figures indicates that the relationships are not entirely straightforward. For example, samples that were 90-100% intelligible received acceptability ratings that ranged across half of the scale (> 5 cm). Similarly, sample that were between 50 and 75% intelligible received PLE ratings that ranged

across more than half the scale (> 6 cm). These wide ranges of ratings suggest that for any given sample, "real world" associations between perceptual dimensions and intelligibility scores may be much lower. That is, for any given sample, it is difficult to predict an unfamiliar listener's perceived effort in listening to that speaker, based only on that person's intelligibility or acceptability score. Thus, it appears that there is a continued need for multiple methods of speech and voice evaluation, including the use of a measure such as PLE. However, further work using qualitative analysis will clarify whether these differences are indeed perceived by these listeners.

The unique contribution of acceptability to PLE, in the absence of a unique contribution of intelligibility, suggests a stronger relationship with the former than the latter. Ratings of both PLE and acceptability have been found sensitive for differentiating equally highly intelligible speech (Eadie & Doyle, 2005; Eadie, Doyle, Hansen & Beaudin, 2008; Nagle & Eadie, 2012b). To interpret these relationships appropriately, it will be necessary to establish whether and how the sensitivity of the scale varies across the range of these dimensions. It seems likely, for example, that ratings of PLE could be most useful for samples of high or "perfect" intelligibility. Houben, van Doorn-Bierman and Dreschler (2013) found that "listening effort" decreased at very high intelligibility levels when signal-to-noise ratio was increased. They used a dual task paradigm in which listening effort was measured objectively as response time to a secondary task presented during a speech recognition task. Reduced response time on a secondary task is an indication that the primary task (i.e., speech processing) is less taxing and can proceed more quickly than if the conditions were challenging (Fisk, Derrick & Schneider, 1986; Gosselin & Gagné, 2010). This finding indicates that despite having decoded an entire utterance to attain 100% intelligibility, listeners were able to demonstrate more efficiency in doing so. Using this dual-task method of obtaining information about listening effort is not clinically feasible, but

ratings of PLE may provide clinically meaningful differentiation of equally intelligible samples with similar perceptual qualities.

4.1.2 Stimulus type and sample size

Results of this study were consistent with reports from other examinations of relationships between PLE and intelligibility using fewer samples. Previous studies have reported nonlinear relationships between PLE and intelligibility for relatively small data sets (Beukelman et al., 2011; 2014; Whitehill & Wong, 2006) and with no distinction among stimuli as to length. Beukelman et al. (2011, 2014) used stimuli elicited using 5 to 11-word sentences from the *SIT* (Yorkston, Beukelman, Hakel & Dorsey, 2007), a Windows-based computer version of the *AIDS,* which was used in the current study. Whitehill and Wong (2006) used 9 to 11-syllable sentences. The current study used similar sentence stimuli, but analyzed data by sentence length group, either Short (5-7 words) or Long (13-15 words). This finding of a length-based difference in the PLE-intelligibility relationship provides a partial explanation for the nonlinearity found in the earlier studies. Results showing an effect of length on PLE ratings for equivalently intelligible samples support the view that this relationship is not linear.

In the current study, a previously unexamined speech type was investigated to address whether the relationships among perceptual dimensions was dependent on speech type. Findings suggest that these relationships do not necessarily differ for different speech types, given similarities in findings for tracheoesophageal (Nagle & Eadie, 2012b), ADSD (Nagle & Eadie, 2012a), dysarthric (Beukelman et al, 2011; 2014) and now electrolaryngeal speech. The current study also used a much larger group of samples than has been used in previous work, with very similar findings. Overall, results in the current study provide further evidence of the previously

identified nonlinear relationship between PLE and intelligibility and strong and unique contribution of acceptability to PLE.

4.2 Differing effects of sentence length

Listening to disordered speech is a challenging task. The more challenging the task is, the more cognitive load is placed on the listener. A longer sentence produced with disordered speech is likely to require more cognitive resources to process. As the resources normally used to attend, remember or ignore irrelevant stimuli are re-allocated to processing disordered speech, the accuracy and speed of processing are likely to suffer and ratings of PLE are likely to increase. Judgments of acceptability, on the other hand, should be less strongly related to discrete decoding and should be more strongly related to an overall assessment of adequacy (or pleasantness, naturalness, quality) of a sample. Even though intelligibility clearly plays a role in speech acceptability, the purpose of obtaining speech quality ratings is to evaluate speech using a perceptual gold standard. Memory *per se* should not be relevant. The finding that sentence length does not significantly affect ratings of acceptability, but does affect ratings of PLE, corresponds with this view of acceptability as a gestalt measure of speech quality.

It appears that equally intelligible sentences with fewer words require less PLE. This finding may logically extend to less complex utterances with more high-frequency vocabulary and greater predictability, although these variables remain to be tested. No length-based difference was found for ratings of acceptability, however, providing initial evidence that for longer sentences, ratings of PLE may provide unique information about disordered speech sample. The difference in PLE ratings in the absence of a difference in ratings of acceptability for the samples with the same intelligibility indicates a potential clinical role for obtaining

ratings of PLE. When intelligibility scores or other auditory-perceptual ratings fail to differentiate speech samples, PLE may be a sensitive indicator of change or difference.

4.3 Perceived listener effort

If PLE is to be used as an outcome measure, it must provide information beyond what is available from currently used clinical measures. The regression model presented in the current study contains only one factor that uniquely predicts PLE, but together the predictors account for a large amount of variance in PLE scores. The finding that acceptability uniquely predicts PLE ratings, in the absence of such a relationship with intelligibility, is consistent with the results of Nagle and Eadie's (2012a) study of equally highly intelligible speech. The fact that only 12% was unique to PLE is also consistent with the univariate results presented in Table 3.1, which show that dimensions such as PLE, acceptability, and intelligibility are highly related constructs. Yet, the results do support the view that a listener might "decode" speech without finding it acceptable or easy to listen to.

Effects of intelligibility and acceptability on PLE ratings are consistent with a theoretical model of listener effort that predicts reduced effort with increasing SNR, such as Hustad and Weismer's (2007) illustration of factors affecting intelligibility and listener comprehension (p. 269). Such a model predicts that factors related to the speaker, the listener and the environment/context contribute to the perception of listener burden. A clear signal from a speaker, with an unimpaired voice source and articulatory filter, would be expected to reduce PLE, other factors being equal. Even presented in noise, a clear signal should require relatively low listener effort (Desjardins & Doherty, 2013; Zekveld & Kramer, 2014). Similarly, a normal-hearing listener should expend less effort listening to a clear speech signal presented in quiet than in noise. Based on findings regarding the role of working memory in processing speech,

such a model is strengthened by the finding that longer sentences place a heavier burden on listeners than shorter sentences (Rönnberg, Rudner, Foo & Lunner, 2008; Tamati, Gilbert & Pisoni, 2013).

The statistical model presented in the current study tested three possible predictors of PLE and predicted 86% of the variance within the sample set. However, individual listener characteristics such as age, working memory capacity, hearing status, receptive vocabulary and motivation are also likely to have effects on ratings of PLE, as they do on objective measures of listening effort (McAuliffe, Wilding, Rickard & O'Beirne, 2012; Picou & Ricketts, 2014; Rudner, Lunner, Behrens, Thorén & Rönnberg, 2012; Tamati et al., 2013). Including these factors in a statistical model of PLE for individual listeners would account for even more of the variance in ratings. These variables deserve future study as they may also help differentiate sensitivity of the scale within different ranges of intelligibility.

4.4 Limitations & future directions

Although the electronic speech produced for this study demonstrated a full range of intelligibility and scale-spanning perceptual ratings, a potential limitation to the findings is the use of healthy speakers. The intact laryngeal anatomy of the speakers was necessarily different from that of any actual user of an electrolaryngeal device, so it is possible that the speech samples used in the study did not represent what might be heard in the "real world." Similarly, the speakers in this study were relatively inexperienced in the use of the electrolarynx device. Actual alaryngeal users of electrolaryngeal devices often improve their speech with training to emphasize plosive and fricative consonants. Through interactions with unfamiliar listeners they may also learn compensatory strategies that allow them to produce speech that is superior to what was used here. Despite these differences, results provide important information about the

perceptual components of PLE in general, and have implications for alaryngeal speakers in particular. Knowing that listeners may struggle to participate in communicative interactions when utterances are longer, for example, may allow them to compensate by producing fewer words in a given conversational turn or to pause more frequently. This knowledge may also be useful in understanding poor social outcomes for users of electrolaryngeal devices (Carr, Schmidbauer, Majaess & Smith, 2000; Eadie, 2007).

Intra- and inter-rater reliability were not dependent variables in the current study, but maximizing the consistency of listener judgments is critical for studies of auditory-perception. Several steps were taken to increase reliability in this study, including setting strict agreement criteria, familiarizing listeners with the tasks and stimuli, and reminding them of the reversed nature of the scale in the second experiment relative to the first. Despite these efforts, data from more than half of the listener participants were dropped from statistical analysis. The number of subjects dropped from this study reflects the particularly subjective nature of PLE and the inherent characteristic of perceptual ratings to shift with exposure or experience to the type of stimulus being rated.

Some of the variability in listener judgments may have been deliberate. What the listeners reported may have been truly different perceptions of their effort in listening to samples of varying quality. As mentioned earlier, listener factors such as relatively low working memory capacity or receptive vocabulary could have systematically affected PLE ratings. Experience in listening to disordered speech samples over the course of the experiment may also have caused their internal standards to shift (Kreiman et al., 1993). Without anchor samples to "reset" them, listeners were forced to rely on their own internal standard of acceptability and expended effort.

There was some evidence of a task effect as well, based on review of each subject's ratings compared to the average ratings from the group rating each block of sentences. Specifically, the VAS was vertically aligned in order to orient "more" as higher, rather than "rightward" as it would be on most horizontal VAS. The vertical alignment was meant to reduce effects of handedness and to alleviate the confusion caused by rating inversely related dimensions (i.e., more acceptable samples are likely to require less effort). Rating tasks were separated by at least one week to further reduce confusion about the orientation of the scales. Despite these efforts, several listeners reported accidentally rating a sample "backwards," and identification of individual ratings of low effort or high acceptability for sentences that were not intelligible confirmed this behavior. Listeners may also have misunderstood the concept or the task in ways that were not revealed through quantitative data analysis. Yet, these data were likely removed by setting strict reliability criteria, increasing the validity of the overall results.

An essential remaining question about PLE is how the dimension itself is interpreted by everyday listeners, who have provided the ratings used in the research reviewed here. The benefit of objective measures of listening effort, such as reaction time and performance measures, is that they are relatively easy to operationally define. PLE, on the other hand, may be interpreted in different ways by different listeners, and that interpretation may change with exposure to the concept or to a rating scale. Definition and interpretation of the terminology used in studies of PLE is critical (McGarrigle et al., 2014). Ratings of PLE are meant to focus on the interaction between the stimulus and the receiver (i.e., listener). Listeners should focus both on their own effort and the qualities of the stimulus. It is essential to discover whether they are actually directing their focus to themselves as much to the stimulus when rating PLE; if not,

there may be nothing gained by rating PLE that is not captured already in ratings of acceptability or other perceptual qualities of speech.

Unfamiliar listeners are most likely to be affected negatively by the quality of disordered speech. One of the major reasons to pursue the use of PLE as an outcome measure is that, whereas the practiced ear of the expert may serve to make clinically relevant perceptual distinctions, documenting incremental changes and differentiating among disordered speech samples, the ear of the everyday listener may not make such fine distinctions. Particular characteristics of disordered speech such as the robotic quality of electrolaryngeal speech may be disconcerting or distracting to unfamiliar listeners, whether experts find the speech relatively acceptable or not (Nagle, Isetti & Eadie, 2014). Discovering how everyday listeners interpret the concept of PLE and the task of rating it is essential to developing real-world, functional outcomes for speech and voice disorders. Qualitative work will help reveal these differences.

References

ASHA. (2002). Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) ASHA Special Interest Division 3, Voice and Voice Disorders. Retrieved from http://www.asha.org/uploadedFiles/ASHA/SIG/03/affiliate/CAPE-V-Purpose-Applications.pdf

Baddeley, A. (2000). The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417–423.

Bennett, S., & Weinberg, B. (1973). Acceptability ratings of normal, esophageal, and artificial larynx speech. *Journal of Speech and Hearing Research*, *16*, 608–615.

Beukelman, D., Childes, J., Carrell, T., Funk, T., Ball, L. J., & Pattee, G. L. (2011). Perceived attention allocation of listeners who transcribe the speech of speakers with amyotrophic lateral sclerosis. *Speech Communication*, *53*(6), 801–806.

Beukelman, D., Gillespie, L., Fager, S., & Ullman, C. (2014). Perceived attention allocation of listeners who transcribe the speech of dysarthric speakers with traumatic brain injury. *Journal of Medical Speech-Language Pathology*, *21*(3), 261–266.

Beverly, D., Cannito, M. P., Chorna, L., Wolf, T., Suiter, D. M., & Bene, E. R. (2010). Influence of stimulus sentence characteristics on speech intelligibility scores in hypokinetic dysarthria. *Journal of Medical Speech-Language Pathology*, *18*(4), 9–13.

Boersma, P., & Weenink, D. (2014). Praat: Doing phonetics by computer (Version 5.3.56). Retrieved from http://www.praat.org/

Bradlow, A. R., Torretta, G. M., & Pisoni, D. B. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication*, *20*(3), 255–272. doi:10.1016/S0167-6393(96)00063-5

Bradlow AR, N. L., Pisoni DB. (1999). Effects of talker, rate, and amplitude variation on recognition memory for spoken words. *Perception & Psychophysics*, *61*(2), 206–19.

Bunton, K., Kent, R. D., Duffy, J. R., Rosenbek, J. C., & Kent, J. F. (2007). Listener agreement for auditory-perceptual ratings of dysarthria. *Journal of Speech, Language & Hearing Research*, *50*, 1481–1495.

Chapanis, A., & Gropper, B. A. (1968). The effect of the operator's handedness on some directional stereotypes in control-display relationships. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *10*(4), 303–320. doi:10.1177/001872086801000401

Cote-Reschny, K., & Hodge, M. (2010). Listener effort and response time when transcribing words spoken by children with dysarthria. *Journal of Medical Speech-Language Pathology*, *18*(4), 24–34.

Desjardins, J. L., & Doherty, K. A. (2013). Age-related changes in listening effort for various types of masker noises. *Ear and Hearing*, *34*(3), 261–272. doi:10.1097/AUD.0b013e31826d0ba4

Doyle, P. C. (2005). Clinical procedures for training use of the electronic artificial larynx. In *Contemporary Considerations in the Treament and Rehabilitation of Head and Neck Cancer* (pp. 545–570). Austin, TX: Pro-Ed.

Doyle, P. C., & Eadie, T. . (2005). The perceptual nature of alaryngeal voice and speech. In *Contemporary Considerations in the Treament and Rehabilitation of Head and Neck Cancer: Voice, Speech, and Swallowing* (pp. 113–140). Austin, TX: Pro-Ed.

Eadie, T. L., & Doyle, P. C. (2005). Scaling of voice pleasantness and acceptability in tracheoesophageal speakers. *Journal of Voice*, *19*(3), 373–83. doi:S0892-1997(04)00071-2 [pii] 10.1016/j.jvoice.2004.04.004

Eadie, T. L., Doyle, P. C., Hansen, K., & Beaudin, P. G. (2008). Influence of speaker gender on listener judgments of tracheoesophageal speech. *Journal of Voice*, *22*(1), 43–57. doi:http://dx.doi.org.offcampus.lib.washington.edu/10.1016/j.jvoice.2006.08.008

Eadie, T. L., & Kapsner-Smith, M. (2011). The effect of listener experience and anchors on judgments of dysphonia. *Journal of Speech, Language, and Hearing Research*, *54*(2), 430–447. doi:10.1044/1092-4388(2010/09-0205)

Ellis, L. (1999). Magnitude estimation scaling judgments of speech intelligibility and speech acceptability. *Perceptual and Motor Skills*, *88*(2), 625–630. doi:10.2466/PMS.88.2.625-630

Fairbanks, G. (1960). *Voice and articulation drillbook* (2nd ed.). New York: Harper.

Fisk, A. D., Derrick, W. L., & Schneider, W. (1986). A methodological assessment and evaluation of dual-task paradigms. *Current Psychology*, *5*(4), 315–327. doi:10.1007/BF02686599

Fraser, S., Gagne, J.-P., Alepins, M., & Dubois, P. (2010). Evaluating the effort expended to understand speech in noise using a dual-task paradigm: The effects of providing visual speech cues. *Journal of Speech, Language & Hearing Research*, *53*(1), 18–33.

Globlek, D., Stajner-Katusic, S., Musura, M., Horga, D., & Liker, M. (2004). Comparison of alaryngeal voice and speech. *Logopedics Phoniatrics Vocology*, *29*(2), 87–91.

Gosselin, P., & Gagne, J.-P. (2011). Older adults expend more listening effort than young adults recognizing speech in noise. *Journal of Speech, Language, and Hearing Research*, *54*(3), 944–958.

Gosselin P.A, & Gagne J.-P. (2010). Use of a dual-task paradigm to measure listening effort. *Canadian Journal of Speech-Language Pathology and Audiology*, *34*(1), 43–51.

Hillman, R. E., Walsh, M. J., Wolf, G. T., Fisher, S. G., & Hong, W. K. (1998). Functional outcomes following treatment for advanced laryngeal cancer. Part I--Voice preservation in advanced laryngeal cancer. Part II--Laryngectomy rehabilitation: the state of the art in the VA System. Research Speech-Language Pathologists. Department of Veterans Affairs Laryngeal Cancer Study Group. *Annals of Otology, Rhinology & Laryngology, Supplement*, *172*, 1–27.

Hustad, K. C. (2006). A closer look at transcription intelligibility for speakers with dysarthria: evaluation of scoring paradigms and linguistic errors made by listeners. *American Journal of Speech-Language Pathology*, *15*(3), 268–77. doi:15/3/268 [pii] 10.1044/1058-0360(2006/025)

Hustad, K. C., & Weismer, G. (2007). A continuum of interventions for individuals with dysarthria: Compensatory and rehabilitative treatment approaches. In *Motor speech disorders : essays for Ray Kent* (pp. 261–303). San Diego: Plural Publishing.

Hustad KC, & Cahill MA. (2003). Effects of presentation mode and repeated familiarization on intelligibility of dysarthric speech. *American Journal of Speech-Language Pathology*, *12*(2), 198–208.

Kalb, M. B., & Carpenter, M. A. (1981). Individual speaker influence on relative intelligibility of esophageal speech and artificial larynx speech. *Journal of Speech & Hearing Disorders*, *46*(1), 77–80.

Kempster, G. B., Gerratt, B. R., Verdolini Abbott, K., Barkmeier-Kraemer, J., & Hillman, R. E. (2009). Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol. *American Journal of Speech-Language Pathology*, *18*(2), 124–32.

Kent, R. D. (1996). Hearing and believing: Some limits to the auditory-perceptual assessment of speech and voice disorders. *American Journal of Speech-Language Pathology*, *5*, 7–23.

Kreiman, J., & Gerratt, B. R. (1998). Validity of rating scale measures of voice quality. *Journal of the Acoustical Society of America*, *104*(3 Pt 1), 1598–608.

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

Larsby, B., Hallgren, M., Lyxell, B., & Arlinger, S. (2005). Cognitive performance and perceived effort in speech processing tasks: Effects of different noise backgrounds in normal-hearing and hearing-impaired subjects. *International Journal of Audiology*, *44*(3), 131–143.

Law, I., Ma, E. P.-M., & Yiu, E. M.-L. (2009). Speech intelligibility, acceptability, and communication-related quality of life in Chinese alaryngeal speakers. *Archives of Otolaryngology - Head and Neck Surgery*, *135*(7), 704–711.

McAuliffe, M. J., Wilding, P. J., Rickard, N. A., & O'Beirne, G. A. (2012). Effect of speaker age on speech recognition and perceived listening effort in older adults with hearing loss. *Journal of Speech, Language, and Hearing Research*, *55*, 838–847. doi:10.1044/1092-4388(2011/11-0101)

McGarrigle, R., Munro, K. J., Dawes, P., Stewart, A. J., Moore, D. R., Barry, J. G., & Amitay, S. (2014). Listening effort and fatigue: What exactly are we measuring? A British Society of Audiology Cognition in Hearing Special Interest Group "white paper." *International Journal of Audiology*, 1–13. doi:10.3109/14992027.2014.890296

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30–46. doi:10.1037/1082-989X.1.1.30

Meltzner, G. S., & Hillman, R. E. (2005). Impact of aberrant acoustic properties on the perception of sound quality in electrolarynx speech. *Journal of Speech, Language & Hearing Research*, *48*(4), 766–79.

Most, T., Tobin, Y., & Mimran, R. (2000). Acoustic and perceptual characteristics of esophageal and tracheoesophageal speech production. *Journal of Communication Disorders*, *33*, 165–181.

Nagle, K., & Eadie, T. (2012a). *Listener effort as an outcome measure for adductor spasmodic dysphonia*. Poster presented at the ASHA Convention, Atlanta, GA.

Nagle, K. F., & Eadie, T. L. (2012b). Listener effort for highly intelligible tracheoesophageal speech. *Journal of Communication Disorders*, *45*(3), 235–245. doi:10.1016/j.jcomdis.2012.01.001

Nagle, K. F., Eadie, T. L., Wright, D. R., & Sumida, Y. A. (2012). Effect of fundamental frequency on judgments of electrolaryngeal speech. *American Journal of Speech-Language Pathology*, *21*(2), 154–166. doi:10.1044/1058-0360(2012/11-0050)

Nagle, K., Isetti, D., & Eadie, T. (2014, March 1). *Everyday listeners' perceptions of adductor spasmodic dysphonia (ADSD) speech*. Poster presented at the 17th Biennial Conference on Motor Speech, Sarasota, FL.

Panico, J., & Healey, E. (2009). Influence of text type, topic familiarity, and stuttering frequency on listener recall, comprehension, and mental effort. *Journal of Speech, Language & Hearing Research*, *52*(2), 534–546. doi:10.1044/1092-4388(2008/07-0238)

Picou, E. M., & Ricketts, T. A. (2014). Increasing motivation changes subjective reports of listening effort and choice of coping strategy. *International Journal of Audiology*, *0*, 1–9. doi:10.3109/14992027.2014.880814

Picou, E., Ricketts, T., & Hornsby, B. (2011). Visual cues and listening effort: individual variability. *Journal of Speech, Language, and Hearing Research*, *54*(5), 1416–1430.

Pindzola, R., & Cain, B. (1988). Acceptability ratings of tracheoesophageal speech. *Laryngoscope*, *98*(4), 394–7.

Portney, L., & Watkins, M. (2000). *Foundations of clinical research : Applications to practice* (2nd ed.). Upper Saddle River: Prentice Hall Health.

Rönnberg, J., Rudner, M., Foo, C., & Lunner, T. (2008). Cognition counts: a working memory system for ease of language understanding (ELU). *International Journal of Audiology*, *47*(S2), S99–S105.

Rudner, M., Lunner, T., Behrens, T., Thorén, E. S., & Rönnberg, J. (2012). Working memory capacity may influence perceived effort during aided speech recognition in noise. *Journal of the American Academy of Audiology*, *23*(8), 577–589. doi:10.3766/jaaa.23.7.7

Sentence Intelligibility Test for Windows. (1996). Lincoln, NE: Institute for Rehabilitation Science and Engineering at Madonna Rehabilitation Hospital.

Sentence Intelligibility Test for Windows. (2007). Lincoln, NE: Institute for Rehabilitation Science and Engineering at Madonna Rehabilitation Hospital.

Southwood, M. H., & Weismer, G. (1993). Listener judgments of the bizarreness, acceptability, naturalness, and normalcy of the dysarthria associated with amyotrophic lateral sclerosis. *Journal of Medical Speech-Language Pathology*, *1*(3), 151–161.

Stewart, C. F., Allen, E. L., Tureen, P., Diamond, B. E., Blitzer, A., & Brin, M. F. (1997). Adductor spasmodic dysphonia: standard evaluation of symptoms and severity. *Journal of Voice*, *11*(1), 95–103. doi:10.1016/S0892-1997(97)80029-X

Sussman, J. E., & Tjaden, K. (2012). Perceptual measures of speech from individuals with Parkinson's disease and multiple sclerosis: Intelligibility and beyond. *Journal of Speech, Language & Hearing Research*, *55*(4), 1208–1219. doi:10.1044/1092-4388(2011/11-0048)

Tamati, T. N., Gilbert, J. L., & Pisoni, D. B. (2013). Some factors underlying individual differences in speech recognition on PRESTO: A first report. *Journal of the American Academy of Audiology*, *24*(7), 616–634. doi:10.3766/jaaa.24.7.10

Tun, P. A., McCoy, S., & Wingfield, A. (2009). Aging, hearing acuity, and the attentional costs of effortful listening. *Psychology and Aging*, *24*(3), 761–766. doi:10.1037/a0014802

Weiss, M. S., & Basili, A. G. (1985). Electrolaryngeal speech produced by laryngectomized subjects: perceptual characteristics. *Journal of Speech & Hearing Research*, *28*(2), 294–300.

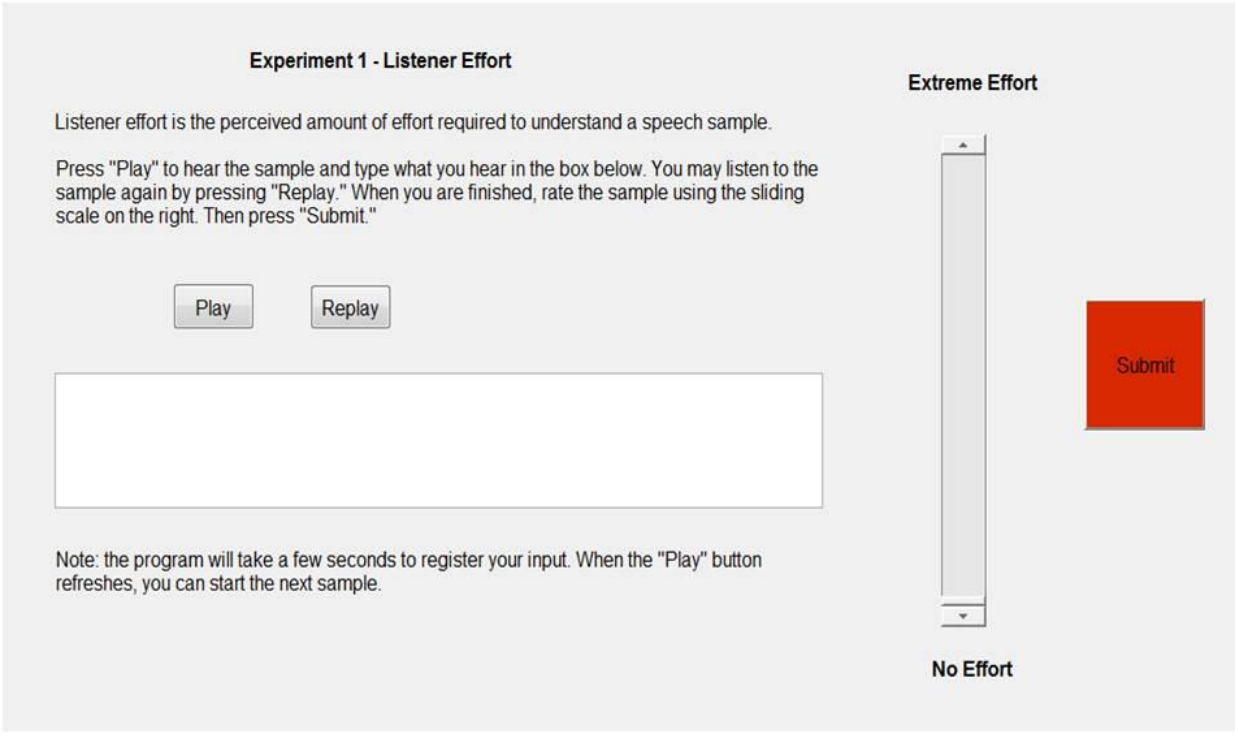
Whitehill, T., & Wong, C. (2006). Contributing factors to listener effort for dysarthric speech. *Journal of Medical Speech-Language Pathology*, *14*, 335–342.

Yorkston, K., & Beukelman, D. R. (1984). *Assessment of intelligibility of dysarthric speech*. Austin TX: Pro-Ed.

Zekveld, A. A., & Kramer, S. E. (2014). Cognitive processing load across a wide range of listening conditions: Insights from pupillometry. *Psychophysiology*, *51*(3), 277–284. doi:10.1111/psyp.12151

Appendix

Initial instructions for transcription & rating task (perceived listener effort)



Chapter 4
Everyday listener impressions of perceived effort when listening to electrolaryngeal speech

Abstract

Everyday listeners transcribed and rated electrolaryngeal speech samples for perceived listener effort (PLE) and acceptability. They were then interviewed about their impressions of performing the tasks, particularly their interpretations of the perceptual concepts being rated. These listeners interpreted acceptability and PLE as strongly negatively associated concepts with a mutual component of understandability. They inferred an additional component of pleasantness or "tolerability" in rating acceptability, whereas their interpretations of PLE had much more to do with the burden/ease of listening to disordered speech. Listeners reported using different strategies to rate samples than to transcribe them, which, along with descriptions of the second listening session as being easier than the first, suggests that listening for comprehension and listening for overall comparison are different processes. Both listening to and understanding electrolaryngeal speech were difficult for these listeners. The social penalties of using electrolaryngeal speech have clinical implications for alaryngeal speakers and individuals with other communication disorders. In addition to exploring more about how listeners perceive disordered speech, future research should investigate technological and compensatory strategies for improving electrolaryngeal speech.

1. Introduction

Hearing loss in the peripheral or central auditory systems has repeatedly been shown to increase the amount of effort required for listeners to process speech (Desjardins & Doherty, 2013; Larsby, Hallgren, Lyxell & Arlinger, 2005), but listening difficulty is not limited to hearing-impaired individuals. Listener characteristics other than hearing status, such as age, working memory capacity and receptive vocabulary, also affect objective measures of listening effort (George, Zekveld, Kramer, Goverts, Festen & Houtgast, 2007; Hallgren, Larsby, Lyxell & Arlinger, 2005; Tamati, Gilbert & Pisoni, 2013). Features external to the listener, such as noise type, noise/presentation level and visual cues, have been shown to affect objective measures of listening effort as well (Brungart, Iyer, Thompson, Simpson, Gordon-Salant, Shurman et al., 2013; Desjardins & Doherty, 2013; Koelewijn, Zekveld, Festen & Kramer, 2012; Zekveld & Kramer, 2014). However, the effect on listening effort of a naturally degraded signal – specifically, disordered speech – has not been well-studied.

For individuals with communication disorders, the burden of listening to disordered speech is of critical interest. Traditional speech and voice measures focus on the severity of the disordered speech itself, using physiologic and acoustic measures. Measures obtained from the speaker with the communication disorder, such as self-reported quality of life or communicative participation, also tend to focus on the well-being of the individual with the communication disorder, and do not traditionally include the perspective of communication partners (i.e., listeners).Yet, it matters how much the voice or speech deviates from what is expected, and for this reason, perceptual measures are the gold standard of evaluation of disordered speech and voice (Kent, 1996; Oates, 2009). However, even standard perceptual measures of voice and speech such as a clinician's or a lay listener's judgments about the roughness of the voice,

speech naturalness, acceptability, or intelligibility may not strongly predict an individual's communicative success. The amount of effort required for a listener to process and understand a given individual's speech is bound to affect communicative interactions with that individual. Whether perceived listener effort (PLE) is a unique construct that is not already captured in other perceptual measures of speech and voice remains to be determined, and is one focus of this study.

1.1 Perceived listener effort and other auditory-perceptual outcome measures

As might be expected, speech intelligibility and PLE are strongly negatively correlated; as intelligibility increases, the effort required to understand it tends to decrease. This association has been found for dysarthric speech (Beukelman, Childes, Carrell, Funk, Ball & Pattee, 2011; Beukelman, Gillespie, Fager, & Ullman, 2014; Whitehill & Wong, 2006) and electrolaryngeal speech (Chapter 3). However, Whitehill and Wong (2006) reported that three of their 33 speakers required a high amount of effort despite high mean intelligibility scores. Beukelman et al. (2011) found a nonlinear relationship between attention allocation and intelligibility for their speakers as well; in fact, attention allocation (i.e., effort) peaked for speakers who were at 75-80% mean intelligibility, and dropped for those with both very high and moderate to low intelligibility. Collectively, these findings suggest that PLE may provide information that goes beyond measures of intelligibility.

Three studies have reported that other auditory-perceptual measures correlate strongly with PLE. First, using samples from speakers with adductor spasmodic dysphonia (ADSD), Nagle and Eadie (2012b) reported a strong positive relationship between PLE and overall speech severity (r = .98). Next, Nagle and Eadie (2012a) investigated 20 inexperienced listeners' perceptions of tracheoesophageal speech, a type of alaryngeal speech. They found a strong,

negative relationship between PLE and ratings of speech acceptability (r = -.99). Finally, using samples of electrolaryngeal speech, also a method of alaryngeal speech, Nagle and Eadie reported strong correlations among PLE and acceptability (r = -.88) and intelligibility (r = .83) in Chapter 3 of this document. With such strong correlations with established auditory-perceptual outcome measures, the meaningfulness of adding PLE to a speech or voice evaluation protocol or as a unique outcome measure is not immediately clear.

By examining individual listener data from a number of studies, however, some patterns about PLE emerge. For example, Nagle and Eadie (2012a; 2012b) found differences between ratings of PLE and other perceptual measures when listeners judged speech samples using a paired comparison paradigm. They reported that some equivalently intelligible samples of ADSD speech were rated as both less severe and requiring more effort when the same sample was compared. In other words, listeners perceived exerting more effort for some samples that they considered the less severe of a pair. This phenomenon was noted for tracheoesophageal speech as well; nearly all listeners in their study rated a sample as both requiring more effort and as being more acceptable than the other sample in at least one pair (Nagle & Eadie, 2012a).

While the individual listener results from both of these studies (Nagle & Eadie 2012a; 2012b) reveal some differences in PLE with other measures, one factor that could have impacted the results could relate to the methodology used for obtaining the perceptual ratings. That is, intelligibility was somewhat controlled in these studies because the only elicitation stimulus was the second sentence of the Rainbow Passage (Fairbanks, 1960); listeners did not actually have to understand or "decode" the samples.

To address these possible limitations, Nagle and Eadie tested the relationships among intelligibility, acceptability and PLE for speech samples obtained from healthy speakers using an

electrolarynx (as described in Chapter 3 of this document). Use of electrolaryngeal samples was deemed appropriate for this initial investigation about PLE a) because of the wide range of intelligibility found across speech samples; and b) because the electrolarynx remains a viable and popular alaryngeal speech method for those who have undergone total laryngectomies (Hillman, Walsh, Wolf, Fisher, & Hong, 1998). In addition, a comparison of PLE with speech acceptability was important because speech acceptability is frequently used as an outcome measure for alaryngeal speech (Bennett & Weinberg, 1973; Eadie & Doyle, 2005). Electrolaryngeal speech, in particular, has been described as being reduced in speech intelligibility, sounding monotone, sounding unnatural or mechanical, sounding too quiet, and being inconvenient to use (Meltzner, Hillman, Heaton, Houston, Kober, & Qi, 2005). Listeners also tend to judge electrolaryngeal speech as significantly less acceptable than other methods of alaryngeal speech, as well as typical speech (Bennett & Weinberg, 1973). In other words, findings about PLE and speech acceptability might not only reveal more about these constructs, but they could have clinical implications for those who use electrolarynges as their primary speech method.

In this study (Chapter 3), a range of intelligibility was ensured using elicitation stimuli that consisted of low-predictability Short (5-7 words) and Long (13-15 words) sentences from the *Assessment of Intelligibility for Dysarthric Speech* (AIDS; Yorkston & Beukelman, 1984). Raters transcribed the sentences and rated their PLE and acceptability using a visual analog scale (VAS), with rating sessions at least a week apart. Although PLE was again found to strongly correlate with acceptability (r = -.88), there was an effect of sentence length on PLE ratings that was not found for ratings of acceptability. Specifically, there was a significant interaction between intelligibility and sentence length; as intelligibility increased, ratings of PLE tended to decrease, but this effect was stronger for Short samples (p = .01). The authors interpreted this

difference as related to how listeners apply the two perceptual terms; whereas the acceptability of a sample might be a result of a gestalt judgment, the amount of effort needed to understand it increases with number of words in the sample. It seems likely, for example, that working memory capacity of listeners would have a much smaller effect on ratings of acceptability than it would on PLE ratings. Yet, without qualitative data from the listeners themselves, it is unknown how listeners interpreted these dimensions or what perceptual strategies they use when making their judgments.

1.2 Everyday listeners

The chapters within this document trace a program of research on the effects of listening to disordered speech and voice for individuals lacking experience with such speech. In quantitative research, these individuals have been referred to as "unfamiliar" (Isetti, Xuereb & Eadie, 2014), "inexperienced" (Eadie & Baylor, 2006) or "naive" (van As, Koopmans-van Beinum, Pols & Hilgers, 2003) listeners. In this chapter, they are referred to as "everyday listeners" to highlight both their unfamiliarity with given speech types and speakers, and their role specifically as listeners in a communicative interaction (Klasner & Yorkston, 2002; 2005). Everyday listeners are routinely used in perceptual research because they gauge the impact of a communication disorder on typical interactions as compared to expert speech-language pathologist listeners. An additional benefit to using everyday listeners is that they may exhibit more similar perceptual strategies as a group; their perceptions of a speech sample *per se* have not been influenced by training or experience with a given communication disorder (Kreiman & Gerratt, 1998).

It is worth noting that intra-rater agreement and inter-rater reliability were not strong for all listeners who made ratings of electrolaryngeal speech using visual analog scales in the study

described in Chapter 3. Unlike the previous studies which used paired comparisons, listeners had no external standard against which to judge a given sample. The numerous personal listener factors that can contribute to PLE prohibit the use of anchor samples in rating tasks; there is no "expert" rating against which to compare ratings of PLE. Ratings may reveal something about the interaction between the listener and the speaker, which may be critically important for specific communicative dyads.

Everyday listeners are commonly used in research into communication disorders, yet they may not share a common interpretation of the perceptual dimensions to be rated. For example, in debriefing interviews, listeners have reported interpreting some perceptual constructs differently from what was expected, despite the use of long-standing instructions for rating those constructs (e.g., Bennett & Weinberg's [1973] instructions for rating speech acceptability; Nagle & Eadie, 2012a). It has not been established what these listeners think they are rating when they provide perceptual judgments of either speech quality or of their own effort. Insight into these perceptions would aid our understanding of perceptual constructs, including PLE.

1.3 Research question

Although the results of previous studies are valuable for measuring speech outcomes, they do not answer the question about why or how listeners come to make these judgments. These studies are also limited by circumscribed dimensions and rating scales selected by clinicians and/or researchers. One approach that may provide insight into the basis of perceptual judgments is to complement quantitative findings with qualitative data, taking a mixed research methods approach (Lasch, Marquis, Vigneux, Abetz, Arnould, Bayliss et al., 2010; Onwuebugzie, Bustamante & Nelson, 2010). The important themes revealed by participants can be revealed through qualitative methods of data analysis, accounting for contextual variables and

providing insight into why participants behave the way they do. Systematically collecting and analyzing authentic communicative activities in context also provides clear links between research and clinical practice (Damico & Simmons, 2003).

One method of eliciting qualitative data is the cognitive interview, in which participants are asked to "mentally reinstate" the physical and personal contexts existing at the time of an experience (Memon & Bull, 2006). Responses are transcribed and coded to extract common themes. For example, pilot data obtained concurrently with Nagle and Eadie (2102a) reflected thematic differences between how listeners interpreted speech acceptability and listener effort for equally intelligible tracheoesophageal speakers. Listeners described how they rated acceptability in terms of the signal itself, comparing it to some standard, and made references to the vocal effort of the speaker (i.e., strained sounding "voice"). Conversely, they described rating PLE not by comparing samples to an internal or external standard, but in terms of the effort required of them to listen. Thus, PLE and acceptability seemed to represent different constructs despite the strong, negative correlation between ratings.

In the current study, a group of everyday listeners was asked to reflect on the task of transcribing and rating PLE and acceptability for electrolaryngeal speech described in Chapter 3 of this document. They provided their impressions of the task of rating these two auditory-perceptual qualities and the specific characteristics of the speech samples that affected their ratings. Qualitative data was obtained to verify that everyday listeners interpret and apply the terms "listener effort" and "acceptability" as intended. Ultimately, in combination with quantitative findings, results should allow refinement of task instructions and definitions of these perceptual qualities, which are currently rated by experts in clinic and everyday listeners in research. The broad question related to the interpretation of PLE and speech acceptability was:

1. How do inexperienced listeners interpret and apply the terms PLE and speech acceptability when rating samples of electrolaryngeal speech?

2. Methods

2.1 Listener participants

Twenty-five everyday listeners (10 male) were interviewed following the experiments described in Chapter 3. Briefly, participants transcribed speech samples provided by healthy male speakers using a monopitch Solatone™ electrolarynx (Griffin Laboratories, Temecula, CA) set at 75Hz. They then rated the speech samples for acceptability or PLE using 100 mm VAS. The following instructions were used for acceptability:

> "Give careful consideration to the attributes of pitch, rate, understandability, and voice quality. In other words, is the voice acceptable to listen to as a listener? (Bennett & Weinberg, 1973, p. 610). Acceptability can be thought of as naturalness, pleasantness and the degree to which the voice is not distracting."

In contrast, PLE was defined as the amount of effort needed to understand a speech sample (Whitehill & Wong, 2006). Participants returned within three weeks to rate the remaining perceptual dimension. After each rating session, they participated in semi-structured cognitive interviews with the first author. The first session took approximately one hour; because participants did not have to transcribe stimuli in the second session, it took about 30 minutes.

When analyzing qualitative data, the number of participants is not necessarily set *a priori*; rather, data are reviewed until conceptual saturation is achieved. Saturation refers to the point at which further data adds little to any patterns that have already been identified. Although every set of data is different, Guest, Bounce and Johnson (2006) found that the ideal number of interviews falls between six and twelve for most data sets. Given the number of listener

participants in this study, conceptual saturation was expected to be easily reached. Ultimately, data from 14 of the participants (7 male) were fully analyzed before conceptual saturation was achieved. Demographic information about the participants whose data was fully analyzed is provided in Table 4.1.

Table 4.1. Participants' demographic information, with perceptual dimension rated in first session
(PLE = perceived listener effort, ACC = acceptability).

Participant	Age	Sex	Race/Ethnicity	Dimension rated first
GB1	29	F	White	PLE
GB2	23	M	White	PLE
GB4	29	M	White	PLE
GB5	20	F	White	PLE
GB6	20	F	African-American	ACC
GB7	19	M	White	ACC
GB8	19	M	White	ACC
GB9	35	M	African-American	PLE
GB10	33	F	White	PLE
GB11	20	F	White	PLE
GB12	28	F	Pacific Islander	PLE
GB13	19	F	White	PLE
GB14	19	M	White	ACC

2.2 Interview procedure

Participants were asked to describe both the task of rating the samples for PLE and speech acceptability, and what the attributes meant to them in the context of rating the samples. A short set of prompt questions was used to initiate a conversation with the first author, who followed up with additional questions arising from participant responses (Baylor, Burns, Eadie, Britton & Yorkston, 2011; Damico & Simmons-Mackie, 2003). For example, if a participant talked about the "tone" of a sample, the interviewer followed up with a clarification question. If an open-ended question (e.g., "What did you mean by 'tone?'") failed to elicit an answer, the interviewer asked whether "tone" referred to pitch, volume or something else. Sample prompt

questions included, but were not limited to, the following:

1. In your own words, how would you define "listener effort/acceptability?
2. What would you say made a sample require more or less effort/be more or less acceptable?
3. Did you use the whole scale when you were rating the samples?

After the second experimental session, participants were asked whether they thought the perceptual dimensions of PLE and acceptability differed and how. Specifically, they were asked how they might differentiate the rating tasks for other inexperienced listeners. The qualitative data obtained from these interviews were analyzed for thematic similarities and differences among participant descriptions of the task and dimensions.

2.3 Data analysis

All cognitive interviews were transcribed for coding, but transcripts were coded only until conceptual saturation was reached. Data were analyzed using the method of constant comparison (Glaser, 1965). This method involves continuous review of similarities and differences in the data, the patterns identified in the data, and the themes ultimately developed from the data (Sandelowski & Boshamer, 2011). The first author developed a set of codes by initially reviewing transcripts iteratively from nine participants. Review of three additional sets of transcripts provided minor additions to the data, and two more sets of transcripts were reviewed to confirm saturation. No new thematic content was obtained from three of the last five sets of transcripts.

Codes were then used to categorize statements made by the participants based on the patterns that emerged from multiple reviews of the transcripts. To capture the context of participants' comments, "meaning units" as short as a single phrase to as long as several

sentences were eligible for coding. Transcripts were independently coded for each dimension by the first and second authors and a graduate assistant. Discrepancies in coding were discussed until agreement was reached, as per established protocols (Coffey & Atkinson, 1996; Lincoln & Guba, 1985). From these codes, primary and secondary themes were identified by the first author, and consensus was obtained from the other coders as to their suitability. Finally, quotations from the data corpus were chosen to represent common themes in participants' descriptions of rating PLE and acceptability.

In qualitative research, the concepts of credibility, dependability, confirmability and transferability roughly approximate the familiar quantitative concepts of validity, reliability, objectivity and generalizability (Lincoln & Guba, 1985). Credibility represents the degree to which interpretation of the data represents the experiences of the participants. It can be established by choosing suitable participants, citing quotations representative of the data corpus and selecting appropriate "meaning units" for analysis. Researchers, experts and participants should reach agreement on these criteria. In the current study, credibility was established in five specific ways (Hammell, Carpenter & Dyck, 2000; Lincoln & Guba, 1985). First, participants included only listeners with little to no experience with or exposure to disordered voice. Second, responses were clarified with participants at the time of each interview. Third, the entirety of each participant's response to the interview questions was examined for both specific and general references to speech acceptability. Fourth, concepts were independently coded for similar themes by all of the authors. Discrepancies in coding were discussed until agreement was reached (Coffey & Atkinson, 1996; Lincoln & Guba, 1985). Finally, quotations from the data corpus were chosen to represent common themes in participants' descriptions of rating overall severity.

Some of the methods used to establish credibility also provide evidence of dependability and confirmability. For example, obtaining clarification and verification from participants about the content of an interview provides some evidence of dependability. Confirmability in the current study was addressed by using multiple coders and working toward agreement among a team with doctoral training in qualitative methods. A third party may also review the acceptability of both the process and the product of qualitative research, in what is termed an inquiry audit (Lincoln & Guba, 1985). For the current study, raw data, observation notes and forms were retained in all iterations in case of an audit.

Transferability refers to the extent to which findings can be generalized to other settings or groups; however, it is the nature of qualitative data to apply specifically to the context in which they were obtained. Judgments about the transferability of qualitative data can only be made by those desiring to apply it. Thus, transferability is addressed through thick description of a situation and its context; interpretation of applicability to a different population is left to the consumer of the research. For this study, to the extent that observations can be recorded during cognitive interviews, thick descriptions of the participants' actions and characteristics were obtained and reported.

3. Results

Four broad coding categories were developed after multiple reviews of the interview transcripts. Although the perceptual qualities of electrolaryngeal speech were not a focus of this study in particular, they were sometimes mentioned by participants as having an effect on their ratings of PLE or acceptability. Thus, the first category (descriptions of the speech signal) had nine associated codes encompassing comments related to noise, articulation, rate, prosody, pitch, sentence length and multidimensional features such as naturalness, pleasantness and clarity.

Impressions of the dimensions of PLE and acceptability were coded separately, and included similarities and differences in how they were interpreted. The fourth category of codes was developed for comments related to the task of listening and rating the samples, such as strategies used to comprehend speech, use of the rating scale and emotional components of performing the tasks.

As a result of iterative review of the coded interviews, three major themes emerged, with seven sub-themes. These themes are represented in Table 4.2 and described further in the following sections.

Table 4.2. Themes and sub-themes derived from content analysis of qualitative data.

Primary Themes	**Sub-Themes**	**Examples**
Everyday listeners interpreted perceptual dimensions as intended	Acceptability and PLE are closely related	understanding speech attributes have similar (if inverse) effects
	There are differences between acceptability and PLE for some listeners	pleasantness (acceptability) vs. ease (PLE)
Everyday listeners described the increased burden associated with listening to electrolaryngeal speech	Electrolaryngeal speech is difficult to listen to	annoying painful tolerated
	Listening to electrolaryngeal has an affective component	frustrating had to really think painstaking
	The second rating session was easier than the first	experience: "I felt more confident" "recognized the sentences"
		no transcription in second session "low level anxiety" in first session
Everyday listeners took different approaches to rating speech stimuli versus comprehending them.	Listeners used top-down strategies to comprehend & transcribe	memorize fill in the blanks guess based on vocabulary
	Listeners made comparisons to other speech when rating samples	internal standards for acceptability only use of scale

3.1 Theme 1: Interpretations of perceived listener effort and acceptability

The first major theme was: *Everyday listeners interpreted perceptual dimensions as intended*. One of the goals of this study was to determine how everyday listeners interpreted the task of rating speech stimuli within and across the two perceptual dimensions.

3.1.1 Sub-Theme: Acceptability and PLE are closely related

There were numerous conceptual similarities between PLE and acceptability. Nearly all of the participants mentioned that ratings of both acceptability and PLE depended on whether they understood the speech sample. When asked about acceptability, many participants

mentioned understandability:

> *I'm definitely putting a lot of effort into all of them but the acceptability kind of more translated as understandability to me. [GB2]*
>
> *What I determined was very acceptable's can I understand it. [GB4]*
>
> *Perceived listener effort wasn't too different for me, my acceptability was definitely like whether or not I felt like I could understand it, if it was something that I had to really really think about then obviously I thought that it was low acceptability. [GB5]*

When describing PLE, participants also talked about understandability, but usually in terms of the amount of effort it took them to "decode" the sentence. For example, Listener GB2 said, "I was rating a very high amount of effort for the listener because I just could not understand." Other participants appeared to agree with this interpretation of PLE:

> *This time I was rating my effort in which I used to understand what the person was trying to say. [GB6]*
>
> *It was the amount that I felt like I had to really think to understand what they were saying. [GB5]*
>
> *I was focused on how easy or difficult it was for me to hear it. [GB7]*
>
> *You have to use maximum effort to understand. [GB9]*

Some participants also thought of PLE and acceptability as two sides of the same coin; the more acceptable the speech was, the less effort it required to understand. For example, Listener GB6 said, "(ACC & LE were) exactly reversed, exactly." Listener GB4, however, spoke of both understanding and working hard (the hallmark of PLE) when rating acceptability: "Acceptability to me is I guess how difficult it is to understand a person and how hard you have to work."

3.1.1.1 Similarities: Speech attributes

Acceptability and PLE were also related in terms of the attributes of the speech samples that affected perceptual ratings; in other words, there was no one attribute that was said to affect

PLE ratings differently from acceptability. As Listener GB 1 stated, “None of them are natural or pleasant.” Participants mentioned the radiated noise from the electrolarynx as decreasing acceptability and increasing PLE, particularly when it was “louder than the actual voice.” [GB1] Listener GB10 echoed this impression, saying “It seemed like when the noise of the apparatus itself was louder than the words being produced that made it difficult to interpret.” Several participants described the synthetic sound of electrolaryngeal speech as “really electronic” [GB8], and “kind of staticky” [GB5]. For example, Listener GB1 said, “It sounds very sort of mechanical, like a robot, and distorted.”

A few speakers had produced stimuli with aspirated stop consonants and fricatives with high frequency noise, and participants noted that these overarticulated productions were easier to understand and more acceptable:

> *If I could tell where the words ended like the ones with the “p” was really strong, I understand. [GB1]*
>
> *Some were able to make a very clear sound and it was very easy to hear their consonants. [GB2]*
>
> *You could tell that some people had to really work on like their “p”s.... [GB8]*
>
> *When there was emphasis placed on “t”s and “p”s, that that definitely improved my comprehension. [GB10]*

The high frequency noise of overemphasized fricatives and aspiration were interpreted by some participants as “whispering:”

> *I was used to listening for the sort of the whisper underneath it cause that’s easier for me to understand. The whispers there I listened to that and then I can block out somehow the background noise but otherwise they’re competing. [GB1]*
>
> *I found that it was easier to hear them when you could kinda hear that whisper. [GB8]*

Similarly, productions that were not overarticulated required more effort and were less acceptable:

> *If the sentence that they were reading had none of those indicator letters with the strong "t" or the "p" or the "s" as a clue to what those words were it was difficult. [GB10]*

Listener GB2 summed up the comments about overarticulation in this way:

> *It feels like the more effort somebody put into it, instead of making it sound like just common everyday speech between a couple of people, more effort somebody puts into it, the more … what you put in is what you get out.*

Participants also mentioned unusual pauses and fast rate as affecting ratings negatively:

> *the slower ones were the ones that were probably easier, um but they didn't really have pauses in them, so I noticed that made it harder for me, but I think there are like certain places in sentences where we pause. [GB1]*
>
> *Sometimes, it didn't seem like there was the right breaks between words. There were certain ones that enunciated really well and those ones that were very slow and I took time to do every single word. I thought it was really, those I thought were very acceptable. [GB7]*
>
> *As long as they speak slowly and you listen, you can understand what they're saying but I noticed when he was speaking fast, it was really difficult for me to make out what he was saying. [GB9]*
>
> *I think some of them people were talking really fast. The people that talk slower, it was a lot easier to understand [GB2]*

The characteristically flat intonation of the monotone electrolarynx affected ratings as well.

Listener GB1 said:

> *all the voices are the same almost. And you don't get any like emotional content, it's just the sentence and it's totally devoid of any emotions or inflections.*

Pitch was mentioned a number of times by participants as affecting perceptual ratings. Specifically, they seemed to find lower pitch (also referred to by participants as tone) to be more favorable:

> *Some of them almost had like a deeper I guess voice to it, those were a lot easier for me to understand. Deep as like if a man has a deeper voice, that was easier for me to understand. There were some that I felt, the pitch was a lot lower and for those ones I could understand it a lot more [GB7]*
>
> *It seemed like lower tones were easier for me to interpret [GB10]*

Finally, the length of the stimuli affected participants' perceptions of how hard they had to work and how acceptable a sample was. Shorter sentences were generally more favorable:

> *When there was shorter sentences it was easier to hear and to figure out cause you weren't trying to figure out something while they were still talking [GB8]*
>
> *If the sentences aren't very very long it makes it a little bit, it makes it easier to listen. [GB9]*
>
> *Some of them were long so I had to replay them because I would forget as I was typing. [GB11]*

3.1.2 Sub-theme: There are some differences between acceptability and PLE for some listeners

Not all participants interpreted the dimensions of acceptability and PLE as mirror images of one another. Specifically, when rating acceptability, these participants talked about the pleasantness of a sample or listener comfort. Listener GB8 drew a sharp contrast between PLE and acceptability:

> *Acceptability was more how pleasant it was to listen to, and PLE didn't really matter how pleasant it was, just how easy it was to hear what they were saying.... Some of them I could hear them clearly but they weren't necessarily pleasant like some of them had like belching in them. That was kinda unpleasant but it was easy to hear.*

After describing acceptability as how difficult it is to understand a person and how hard you have to work, Listener GB4 added "I almost equate it to comfortability." Participants also spoke of considering the length of time they could tolerate listening to such speech, which affected judgments of acceptability more strongly than PLE. For example, Listener GB7 said:

> *I was like trying to decide whether or not I could you could actually listen to this for like a certain amount of time. You know, is the thing acceptable. The effort is lower because you could understand what they're saying but if you had to listen to it for a long period of time, it would suck.*

Finally, when asked directly whether the dimensions seemed the same to her, Listener 12 said, "No, it definitely felt like I was doing something different. It seemed like rating

acceptability was easier because then it was just the quality of what I was listening to."

3.2 Theme 2: Everyday listeners described an increased burden associated with listening to electrolaryngeal speech.

The second major theme had to do with the burden of listening to electrolaryngeal speech and how that burden affected ratings of PLE and acceptability. Most of the participants stated that the speech stimuli in this study were either hard to listen to because of their quality or that they were difficult to understand, presumably for similar reasons.

3.2.1 Sub-theme: Listening to electrolaryngeal speech has an affective component.

This first sub-theme incorporated affective reactions to electrolaryngeal speech. Some participants made comments about how hearing EL speech made them feel, regardless of the task or the dimension being rated. Listener GB7, for example, stated that he tolerated the speech and "didn't like a couple of them." This type of reaction was not uncommon:

> *(Listening) was annoying…. Ugh, very annoying. [GB6]*
>
> *(I thought) was it painful to listen to because there were some that would really ring in your ear afterwards [GB4]*

There was also an affective component to the task of making decisions about the samples (as opposed to simply hearing them) for Listener GB1:

> *I found that I had to remind myself that I didn't need to have an emotional connection to the word "acceptable" and the word "pleasant" Even though those words were triggering me to be emotional. [GB1]*

3.2.2 Sub-theme: Listening to EL speech (for comprehension) takes work.

This sub-theme specifically addressed the effort component of listening for comprehension. For example, Listener GB5 said that he "had to really think to understand." Listener GB2 said that he was "straining my brain" to understand, and others noted physical reactions to attempting to understand the speech as well:

> *Effort to me means very clear speaking, no noise. Extreme effort, I almost squint my eyes to listen hard. [GB11]*
>
> *I was getting more tired as I went. [GB1]*

3.2.3 Sub-theme: The second rating session was easier than the first.

Some participants mentioned that the second ratings session seemed easier than the first. Some said that having had experience with EL speech in the previous session was helpful. For example, Listener GB4 said that he was "prepared to hear what I was going to hear." Others also mentioned being more familiar with electrolaryngeal speech as making the second session easier:

> *Something I was completely unfamiliar with I probably never heard anybody actually speak like this before and now after just one session I have a better understanding. [GB2]*
>
> *The second time doing it so I felt more confident with it, and then it felt less like a test so I was less anxious [GB1]*

Although within a session, experience with speech may have been tempered by fatigue:

> *I noticed getting more tired as it went on, like at the same time, I got used to it, so I got better at it and I got more tired, so it was harder....[GB1]*

Another reason that the second session may have been easier for these participants was because they did not have to transcribe the speech samples. In the words of Listener GB11, in the second session, "there was no pressure to memorize it," presumably for transcription.

> *Rating acceptability was easier because it was just the quality of what I was listening to. Last time, I needed more effort because I really had to focus on what words exactly were said. [GB12]*
>
> *It was easy it was definitely a lot easier than last week because I got to hear everything twice last week and still not have any clue what they were saying which is kind of frustrating, whereas this week it was just listen and did you understand it. [GB2]*

A few participants talked about taking a less active approach in the experimental session that did not require transcription (i.e., the second session). For example, Listener GB12 mentioned that the second session was easier because she "could be a little more passive and it was easier to rate

that way."

3.3 Theme 3: Everyday listeners took different approaches to rating speech stimuli versus comprehending them

The third major theme had to do with differences between the task of transcribing (i.e., comprehending) and the task of rating (i.e., comparing) electrolaryngeal speech for PLE or speech acceptability. The two sub-themes addressed the differing ways in which participants performed the tasks.

3.3.1 Sub-theme: Listeners used top-down strategies to comprehend and transcribe stimuli

The first sub-theme focused on the strategies participants used to "decode" the samples. Numerous strategies were employed to "get the sentence," in the words of Listener GB1. Participants were allowed one repetition of each stimulus in the first session, and many mentioned replaying the samples that were difficult to understand. They also talked about memorizing parts of a stimulus in order to piece it together after hearing it a second time. For example, Listener GB11 said she had to "listen to sentence twice and then memorize them."

"Filling in the blanks" was another top-down strategy used by participants in this study.

> *You can go "maybe this is what he's trying to say" so and then with that I'll kinda like brush off you know the noise. [GB6]*
>
> *If I could pick out a couple of the words paired together, I could assume even if it was kind of staticky I could assume what that person was trying to say. [GB1]*

Referring to rating the stimuli compared to transcribing them, Listener GB1 talked about the difference between having a one-to-one conversation with disordered speech and being in a group.

> *"It's like the difference between being the only person who's talking to someone with a voice box (electrolarynx) and be like 'you're on,' like you have to hold the conversation versus there being five other people and if one of you gets what they're saying, they can save the social situation, so that's kinda how it felt with that difference."*

3.3.2 Sub-theme: Listeners made comparisons to other speech when rating acceptability

The second sub-theme was related to perceptual strategies used by listeners. Participants mentioned comparing the EL samples to "normal" speech or to some internal standard. For example, Listener GB2 said he imagined that a friend was talking to him as he tried to gauge acceptability. Only one speaker (GB7) mentioned comparing the samples to the other samples in the set (as opposed to comparing them to typical speech), which presumably affected how he used the rating scale.

Decisions about the standard against which to judge the samples were critical in how the scale was used by the participants to rate acceptability. When asked, participants said they use "the whole scale," up to "not quite the top," "all but the top two-thirds," and "mostly the middle."

> *I did start to think well if I go all the way up to very acceptable, like what does that mean? Cause none of them are natural or pleasant but I understand what you know, it depends what your standard is. [GB1]*

Although participants mentioned using the scale in similar ways when rating PLE, their judgments were clearly based on their own effort as opposed to comparison of the stimulus with a standard.

4. Discussion

Three main findings emerged from the thematic analysis of data in this study. Everyday listeners interpreted acceptability and PLE as strongly negatively associated concepts with a mutual component of understandability. As a group, participants mentioned common perceptual features of electrolaryngeal speech that affected its acceptability and the amount of effort they required to listen to it. For example, they noted the robotic quality, monotonicity, efforts in articulation, changes in speech rate, etc., that could be detrimental or beneficial to both

acceptability and PLE. However, some participants did recognize a conceptual difference between the two perceptual dimensions examined in this study, with acceptability being associated with pleasantness or comfort, and PLE representing the ease of listening. Second, participants reported that just hearing electrolaryngeal speech was difficult, whether or not they were asked to listen for comprehension (and regardless of the dimension being rated). Third, participant comments suggested that differences between tasks, rather than perceptual dimensions, compelled them to attend to the samples in dissimilar ways. In other words, their strategies for completing the task of rating compared to transcribing the samples varied more than their interpretations of acceptability compared to PLE. However, some notable differences were found in use of the rating scales and perceptual strategies reported by listeners that also lend support towards the differentiation of these two dimensions.

4.1 Perceived listener effort versus acceptability

That participants mentioned understandability and related terms (e.g., comprehensibility, "figuring it out") when providing their impressions of both PLE and acceptability is to be expected. Understandability was mentioned in the instructions for rating both dimensions in this study. Identifying a component of understandability in both dimensions is also consistent with findings of strong positive correlations between acceptability and intelligibility for dysarthric speech (Southwood & Weismer, 1993) and electrolaryngeal speech (Chapter 3; Nagle, Eadie, Wright & Sumida, 2012), and for strong negative correlations between PLE and intelligibility for dysarthric (Beukelman et al., 2011; 2014; Whitehill & Wong, 2006) and electrolaryngeal speech (Chapter 3).

The characteristics of electrolaryngeal speech that increased speech acceptability were also reported to decrease the amount of effort required to understand it. On the whole, slower,

overarticulated short sentences with little radiated noise from the device were more acceptable. This finding is consistent with previous research. Specifically, reduced rate has been linked to improved intelligibility scores for dysarthric speech (Hammen, Yorkston & Minifie, 1994) and improved acceptability for electrolaryngeal speech (Weiss, Yeni-Komshian & Heinz, 1979). Reducing radiated noise (i.e., improving signal-to-noise ratio; SNR) has been shown to improve both intelligibility and acceptability for electrolaryngeal speech (Meltzner & Hillman, 2005). Reports that the more acceptable speech samples requiring less effort contained overarticulated voiceless consonants (i.e., with an aspiration burst or frication noise) is supported by quantitative findings reported for both alaryngeal (Weiss & Basili, 1985) and laryngeal users of electrolarynges (Weiss, Yeni-Komshian, Heinz, 1979).

On the other hand, the study reported in Chapter 3 reported no difference in intelligibility or acceptability for sentences with fewer words (5-7) compared to those with more words (13-15); for PLE, however, shorter sentences required significantly less effort than longer sentences. Perceived listener effort might be said to increase for longer sentences due to the limits of working memory capacity; if, however, acceptability is perceived as a *gestalt*, sentence length should not affect acceptability ratings (Chapter 3). One participant in the current study provided a possible explanation for this when he said that acceptability had to do with how long he could tolerate listening to the samples. This comment reflects the omnipresent link between signal and listener factors in perceptual ratings of speech. Clearly, this participant was describing the effect of a given sample on himself; yet it was the quality of that sample that led him to realize he was "tolerating" it.

Some participants mentioned favoring samples that were produced with relatively lower pitch or "tone." This is consistent with other findings of both increased intelligibility and

acceptability for electronic speech samples with lower average fundamental frequencies (f0) relative to those with higher frequencies (for males; Nagle et al., 2012). However, in this study, all speech samples were produced by males with intact larynges using a monotone device set at 75Hz; actual f0 deviated no more than 2-3 Hz per sample from that setting. The perception of between-sample differences was likely due to variability in the filtering characteristics of individual speakers' vocal tracts.

Several participants did draw distinctions between the concepts of PLE and acceptability. Although highly negatively correlated, these concepts are meant to emphasize different features of the listener-speaker interaction (Lindblom, 1991; Kreiman, Gerratt, Kempster, Erman & Berke, 1993). Specifically, PLE is meant to represent a listener's assessment of his or her own effort in processing speech (McGarrigle, Munro, Dawes, Stewart, Moore, Barry et al., 2014). Acceptability, on the other hand, is meant to capture a listener's observation of not only the quality of a speech sample, but also the comfort or pleasantness of that sample in the context of that listener's standards. In the experiment from which the cognitive interview data were obtained, listeners were provided with instructions that made these distinctions, but the differing emphases of the rating tasks were not otherwise highlighted.

Some participants made casual references to the difference between PLE and acceptability, such as Listener GB4's comment about equating acceptability with "comfortability." This comment followed his description of acceptability as the nearly verbatim definition of PLE provided in the listening experiment, and suggests that he was attending to his own effort or comfort as a listener in both rating tasks. However, several other participants drew sharp distinctions between the dimensions after having rated both. Acceptability had to do with the pleasantness or "quality of what (they were) listening to." They reported that PLE was about

"how hard it was to hear" a given sample. Ultimately, these listeners did seem to make a distinction between the quality of a sample (acceptability) and the ease of listening to it (PLE).

In addition to subtle differences in attributes that contribute to acceptability and PLE, listeners also noted differences in perceptual strategies when performing their ratings. For example, Listener GB2 said he imagined that a friend was talking to him as he tried to gauge acceptability, which was not a strategy used for PLE. This is consistent with pilot data obtained for tracheoesophageal speech from a different group of inexperienced listeners. Listeners reported rating acceptability in terms of the signal itself and comparing it to a standard. PLE was reportedly rated strictly in term of the effort required of them to listen. Results from the present study appear to be consistent with these previous observations for tracheoesophageal speech.

4.3 Reactions to the tasks

Participants reacted in four general ways to the tasks required of them in the experiments preceding these interviews (Chapter 3). First, they talked about the difficulty of understanding the speech samples, regardless of whether they ultimately decoded all of the words. Descriptions of this process as painstaking, painful and tiring underscore the possible utility of measuring PLE; listeners may have to try very hard to understand, even if they are eventually successful. As previously demonstrated by Houben, van Doorn-Bierman and Dreschler (2013) and Munro and Derwing (1995), both objective and subjective measures of listening effort may serve to differentiate equally highly intelligible speech samples.

Participants also described an effect of exposure or practice listening to electrolaryngeal speech. They reported "getting better at listening' and "getting more used to it." More than one participant reported that by the second listening session, he benefited from being able to prepare to hear the speech. The potential benefit of listener familiarity for intelligibility is a well-

described phenomenon, for both disordered speech (e.g., Tjaden & Liss, 1995) and non-native speech (e.g., Gass & Varonis, 1984).

Because of the experimental procedure preceding the cognitive interviews described in this study, it was not possible to determine whether the benefits of familiarity with electrolaryngeal speech extended to improved understanding; participants transcribed samples only in the first listening session. As has been noted in previous research, participants used top-down strategies to decode the disordered speech presented in this study (Klasner & Yorkston, 2002; 2005). Some of their strategies were undoubtedly due to the pressure of having to transcribe the samples (as opposed to what they might have done in an actual conversation). For example, they described memorizing parts of a sample or replaying it in order to understand every word. These participants were very reluctant to give up, at least during the transcription portion of the experiment. On the other hand, feeling *able* to give up listening for comprehension seemed to be a relief.

Participants did describe strategies they would likely use in the real world. For example, they spoke of fitting words together if they were able to hear a few of them in a sentence, and of guessing other words based on the likelihood that they would occur with a word they had understood. In cases where it all became too difficult, they gave up. As Listener GB1 said, there is a difference between "being on" and having to hold a conversation, and being part of a group in which someone else can "save the social situation." This behavior is consistent with the conclusions of Beukelman et al (2011), who interpreted the decline in mean attention allocation scores for samples that were less than 65% intelligible as indicating listeners' realization that they were not comprehending the stimuli. Zekveld and Kramer (2014) also noted a decrease in listening effort (measured by pupil dilation during a challenging listening task) that was related

to increased ratings of "giving up trying."

In the second session, participants were relieved of the pressure of fully comprehending the sample. They had only to rate it for either PLE or acceptability, and the strategies they used were generally different. Participants spoke of relaxing and being more passive about listening than in the first experiment. Some participants mentioned comparing samples to other types of speech, such as "normal speech," "my speech," and other electronic speech. One participant described imagining that it was his friend speaking, presumably to take a more sympathetic approach to rating the samples. Comments such as this are consistent with a mixed methods study of everyday listeners and ADSD speech (Nagle, Eadie & Yorkston, submitted), in which very similar comments and themes were identified. As in the current study, one theme had to do with the difficulty of rating overall severity of samples of ADSD (Nagle et al., submitted). Participants in that study mentioned trying not to compare the ADSD samples to "normal" speech and using the scale as intended. Because the samples in that study were presented in pairs, listeners were provided with an external standard against which to judge each sample (i.e., the other sample in the pair); remembering not to rate samples against "normal" speech appeared to require some effort.

4.4 Reactions to electrolaryngeal speech

Participants talked about how electrolaryngeal speech made them feel as well as the difficulty of actually understanding it. In addition to describing the speech as robotic, mechanical and unemotional, participants spoke of being annoyed by the sound of it. One said that he "just didn't like it." These results are consistent with previous research. For example, listeners in the study by Bennett and Weinberg (1973) stated that the most troublesome characteristic of the electrolarynx used was the "mechanical" nature of the signal produced, and the associated

monotony.

In the current study, difficulty understanding the speech was compounded by distraction by the sound of the speech itself. Some participants mentioned a loudness difference in the favored samples, in which the "voice" was louder than the radiated noise from the device. Preference for a sample with a higher SNR is intuitive and consistent with research into the effects of radiated noise from electrolarynges (e.g., Meltzner & Hillman, 2005) and of masking noise on speech reception thresholds (e.g., Koelewijn, Zekveld, Festen & Kramer, 2012).

Participants in the earlier study also described sympathizing with the speakers who produced the samples of ADSD speech (Nagle et al., submitted). Although not mentioned in the current study, those participants described feeling judgmental about rating the ADSD samples, and that the apparent subjectivity of the task made them uncomfortable. Strictly in terms of frequency, participants rating and describing electrolaryngeal speech were less hesitant to describe their negative reactions overall; their comments about discomfort were less about making judgments and more about the unpleasantness of the sound of electrolaryngeal speech. As reported by Meltzner and Hillman (2005), even with noise reduction and pitch modulation enhancements, electrolaryngeal speech has been judged as far different from normal and synthesize alaryngeal speech. It would seem that the potential social penalties associated with using an electrolarynx are based not only on the difficulty of understanding it, but on its basic sound as well.

4.5 Limitations & Future Directions

The themes revealed in this study suggest that individuals unfamiliar with disordered speech can interpret and follow instructions for auditory-perceptual tasks. Despite this, some participants reported struggling with how to use the rating scale, which was undifferentiated

apart form endpoints. Previous studies of PLE for disordered speech suggested that inexperienced/everyday listeners are highly reliable, but both of these studies obtained ratings through paired comparisons, in which samples were directly compared to one another (Nagle & Eadie, 2012a; Nagle et al., submitted). The use of an unmarked VAS, and the lack of an external standard beyond previous trials, may account for the variety of ways that participants reported using the scale in the current study. Should the use of PLE ratings become more prevalent, it will be necessary to determine what type of rating scale is appropriate to ensure strong reliability.

A few participants commented on the opposing directions of the scales used in the experiments described here. Vertical scales were aligned so that "more" of a given direction was indicated by placing the cursor higher on the scale; "more acceptable" was closer to 10 cm, but "more effort required" was as well. More than one participant described PLE and acceptability as being opposite, and review of both qualitative and quantitative data indicated that some found the layout of the scale confusing. If ratings of PLE are obtained in conjunction with other auditory-perceptual dimensions, the alignment and direction of the rating scale will have to be made very clear.

Several participants remarked that the second task in the experiment was easier than the first. This was not surprising, given that in the first session, they had to both transcribe and rate the sentences for one of the perceptual dimensions; in the second session, they had only to rate the other dimension. Given this methodology, it is not possible to tease out the effects of familiarity or exposure to the samples from the effects of listening for comprehension. Although understandability played a large role in both PLE and acceptability, it is arguably a critical part of PLE. It may be that ratings of PLE should only be obtained in conjunction with transcriptions, or that ratings of acceptability are less meaningful when obtained along with transcriptions. More

research is needed to differentiate the effects of experience with disordered speech from effects of experimental task.

The speech samples used in this study were obtained from healthy male speakers aged 50-65; although this age range was chosen as reflecting the average alaryngeal speaker, all speakers had intact larynges. Results of qualitative research are not often generalizable beyond the population investigated, and in this case, it is possible that actual users of electrolaryngeal speech would encounter different reactions from what is reported here. For example, the nature of electrolaryngeal speech is inherently different from all other types, with electrolaryngeal speech being judged least acceptable among alaryngeal speech types, as well as compared to typical speech (Bennett & Weinberg, 1973). Yet, it is interesting that qualitative data found in this study appear to be consistent with previous quantitative results (strong correlations between acceptability and PLE, but with some individual listener differences; Nagle & Eadie, 2012a) and qualitative findings in other speaker populations (e.g., ADSD speech). How these results generalize to other speakers with disordered voice and speech needs further investigation.

Finally, and most importantly, these results were obtained under controlled conditions in a lab environment. Participants described expending as much effort as necessary and results reported in Chapter 3 suggest that they did. In fact, this level of effort may be unrealistic in a real world situation. For example, Evitts and Searl (2006) examined reaction times for processing different alaryngeal speech types spoken by one proficient alaryngeal speaker. They found that reaction times were highest for the speaker when he used an electrolarynx compared to the other two speech modes. Their results suggest that processing electrolaryngeal speech, in particular, is particularly effortful compared to other modes of speech. Comments about giving up or becoming more passive in the second experimental session suggest that this was also the case in

this study.

Participants in the current study appeared highly motivated, but were not taking conversational turns, as they might have in a real conversation. As Listener GB1 said, the stakes are higher when the listener has to attend and possibly save a social situation singlehandedly. One study that looked at the role of motivation in objective listening effort found that listeners expended more effort in a listening task when told they would be tested than when they were just told to listen carefully (Picou & Ricketts, 2014). Further qualitative research should examine actual conversations between everyday listeners and individuals with disordered speech. Unless conducted over the telephone, a conversation would typically include visual information, which would be expected to reduce listener effort.

5. Conclusion

Qualitative findings support results of previous experiments investigating the use of PLE for disordered voice and speech (Chapter 3; Nagle & Eadie, 2012a; Nagle et al., submitted). Everyday listeners in this study interpreted acceptability and PLE as strongly negatively associated concepts with a mutual component of understandability; however, they described exclusive characteristics of each concept as well. Participants inferred an additional component of pleasantness or "tolerability" in rating acceptability, whereas PLE had much more to do with the burden/ease of listening to disordered speech. Perceptual features of electrolaryngeal speech were identified that affected both acceptability and the amount of effort required to listen to it, suggesting that taking steps to improve the acceptability of speech could have an effect on listening effort. However, caution is warranted in making conclusions about causation, rather than correlation. They reported using different strategies to rate samples than to transcribe them, which, along with descriptions of the second listening session as being easier than the first,

suggests that listening for comprehension and listening for overall comparison are different processes. Both listening to and understanding electrolaryngeal speech were difficult for these listeners. The likely social penalties of using electrolaryngeal speech have clinical implications for alaryngeal speakers, who may use an electrolarynx as a backup speech mode, and potentially for individuals with other communication disorders. In addition to exploring more about how listeners perceive disordered speech, future research should investigate technological and compensatory strategies for improving electrolaryngeal speech.

References

Baylor, C., Burns, M., Eadie, T., Britton, D., & Yorkston, K. (2011). A qualitative study of interference with communicative participation across communication disorders in adults. *American Journal of Speech-Language Pathology*, *20*(4), 269–287. doi:10.1044/1058-0360(2011/10-0084)

Bennett, S., & Weinberg, B. (1973). Acceptability ratings of normal, esophageal, and artificial larynx speech. *Journal of Speech and Hearing Research*, *16*, 608–615.

Beukelman, D., Childes, J., Carrell, T., Funk, T., Ball, L. J., & Pattee, G. L. (2011). Perceived attention allocation of listeners who transcribe the speech of speakers with amyotrophic lateral sclerosis. *Speech Communication*, *53*(6), 801–806.

Beukelman, D., Gillespie, L., Fager, S., & Ullman, C. (2014). Perceived attention allocation of listeners who transcribe the speech of dysarthric speakers with traumatic brain injury. *Journal of Medical Speech-Language Pathology*, *21*(3), 261–266.

Brungart, D., Iyer, N., Thompson, E., Simpson, B. D., Gordon-Salant, S., Shurman, J., … Grant, K. W. (2013). Interactions between listening effort and masker type on the energetic and informational masking of speech stimuli. *Journal of the Acoustical Society of America*, *133*(5), 3435. doi:10.1121/1.4806059

Coffey, A., & Atkinson, P. (1996). *Making sense of qualitative data : complementary research strategies*. Thousand Oaks: Sage Publications.

Damico, J., & Simmons-Mackie, N. (2003). Qualitative research and speech-language pathology: A tutorial for the clinical realm. *American Journal of Speech-Language Pathology*, *12*(2), 131–143.

Desjardins, J. L., & Doherty, K. A. (2013). Age-related changes in listening effort for various types of masker noises. *Ear and Hearing*, *34*(3), 261–272. doi:10.1097/AUD.0b013e31826d0ba4

Gass, S., & Varonis, E. M. (1984). The effect of familiarity on the comprehensibility of non-native speech. *Language Learning*, *34*(1), 65–89.

George, E. L. J., Zekveld, A. A., Kramer, S. E., Goverts, S. T., Festen, J. M., & Houtgast, T. (2007). Auditory and nonauditory factors affecting speech reception in noise by older listeners. *Journal of the Acoustical Society of America*, *121*(4), 2362–2375. doi:10.1121/1.2642072

Glaser, B. G. (1965). The constant comparative method of qualitative analysis. *Social Problems*, *12*(Spring), 436–445.

Guest, G., Bunce, A., & Johnson, L. (2006). How many interviews are enough? An experiment with data saturation and variability. *Field Methods*, *18*(1), 59–82. doi:10.1177/1525822X05279903

Hallgren, M., Larsby, B., Lyxell, B., & Arlinger, S. (2005). Speech understanding in quiet and noise, with and without hearing aids. *International Journal of Audiology*, *44*(10), 574–583.

Hammell, K. W., Carpenter, C., & Dyck, I. (2000). *Using qualitative research : a practical introduction for occupational and physical therapists*. Edinburgh; New York: Churchill Livingstone.

Hammen, V. L., Yorkston, K. M., & Minifie, F. D. (1994). Effects of temporal alterations on speech intelligibility in parkinsonian dysarthria. *Journal of Speech & Hearing Research*, *37*(2), 244–253.

Houben, R., van Doorn-Bierman, M., & Dreschler, W. A. (2013). Using response time to speech as a measure for listening effort. *International Journal of Audiology*, *52*(11), 753–761. doi:10.3109/14992027.2013.832415

Kent, R. D. (1996). Hearing and believing: Some limits to the auditory-perceptual assessment of speech and voice disorders. *American Journal of Speech-Language Pathology*, *5*, 7–23.

Klasner, E., & Yorkston, K. (2005). Speech intelligibility in ALS and HD dysarthria: The everyday listener's perspective. *Journal of Medical Speech-Language Pathology*, *13*(2), 127–140.

Klasner, E., Yorkston, K. M., & Lillvik, M. (2002). Everyday listeners' perspective on Amyotrophic Lateral Sclerosis and Huntington disease dysarthria: barriers and strategies in understanding distorted speech samples. *Journal of Medical Speech-Language Pathology*, *10*, 293–298.

Koelewijn, T., Zekveld, A. A., Festen, J. M., & Kramer, S. E. (2012). Pupil dilation uncovers extra listening effort in the presence of a single-talker masker. *Ear and Hearing*, *33*, 291–300. doi:10.1097/AUD.0b013e3182310019

Kreiman, J., & Gerratt, B. R. (1998). Validity of rating scale measures of voice quality. *Journal of the Acoustical Society of America*, *104*(3 Pt 1), 1598–608.

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

Larsby, B., Hallgren, M., Lyxell, B., & Arlinger, S. (2005). Cognitive performance and perceived effort in speech processing tasks: Effects of different noise backgrounds in normal-hearing and hearing-impaired subjects. *International Journal of Audiology*, *44*(3), 131–143.

Lasch, K. E., Marquis, P., Vigneux, M., Abetz, L., Arnould, B., Bayliss, M., … Rosa, K. (2010). PRO development: rigorous qualitative research as the crucial foundation. *Quality of Life Research*, *19*(8), 1087–1096.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Beverly Hills, Calif.: Sage Publications.

Lindblom, B. (1990). On the communication process: Speaker-listener interaction and the development of speech. *Augmentative & Alternative Communication*, *6*(4), 220–230.

Meltzner, G. S., & Hillman, R. E. (2005). Impact of aberrant acoustic properties on the perception of sound quality in electrolarynx speech. *Journal of Speech, Language & Hearing Research*, *48*(4), 766–79.

Memon, A., & Bull, R. (1991). The cognitive interview: Its origins, empirical support, evaluation and practical implications. *Journal of Community & Applied Social Psychology*, *1*(4), 291–307. doi:10.1002/casp.2450010405

Munro, M. J., & Derwing, T. M. (1995). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language & Speech*, *38*(3), 289–306.

Nagle, K., & Eadie, T. (2012). *Listener effort as an outcome measure for adductor spasmodic dysphonia*. Poster presented at the ASHA Convention, Atlanta, GA.

Nagle, K. F., & Eadie, T. L. (2012). Listener effort for highly intelligible tracheoesophageal speech. *Journal of Communication Disorders*, *45*(3), 235–245. doi:10.1016/j.jcomdis.2012.01.001

Nagle, K. F., Eadie, T. L., Wright, D. R., & Sumida, Y. A. (2012). Effect of fundamental frequency on judgments of electrolaryngeal speech. *American Journal of Speech-Language Pathology*, *21*(2), 154–166. doi:10.1044/1058-0360(2012/11-0050)

Nagle, K., Isetti, D., & Eadie, T. (2014, March 1). *Everyday listeners' perceptions of adductor spasmodic dysphonia (ADSD) speech*. Poster presented at the 17th Biennial Conference on Motor Speech, Sarasota, FL.

Oates, J. (2009). Auditory-perceptual evaluation of disordered voice quality. *Folia Phoniatrica et Logopaedica*, *61*(1), 49–56. doi:10.1159/000200768

Onwuegbuzie, A. J., Bustamante, R. M., & Nelson, J. A. (2010). Mixed research as a tool for developing quantitative instruments. *Journal of Mixed Methods Research*, *4*(1), 56–78. doi:10.1177/1558689809355805

Picou, E. M., & Ricketts, T. A. (2014). Increasing motivation changes subjective reports of listening effort and choice of coping strategy. *International Journal of Audiology*, *0*, 1–9. doi:10.3109/14992027.2014.880814

Sandelowski, M., & Boshamer, C., C. (2011). Parallelism in constant comparison analysis. *Research in Nursing & Health*, *34*(6), 433–434. doi:10.1002/nur.20455

Southwood, M. H., & Weismer, G. (1993). Listener judgments of the bizarreness, acceptability, naturalness, and normalcy of the dysarthria associated with amyotrophic lateral sclerosis. *Journal of Medical Speech-Language Pathology*, *1*(3), 151–161.

Tamati, T. N., Gilbert, J. L., & Pisoni, D. B. (2013). Some factors underlying individual differences in speech recognition on PRESTO: A first report. *Journal of the American Academy of Audiology*, *24*(7), 616–634. doi:10.3766/jaaa.24.7.10

Tjaden, K., & Liss, J. M. (1995). The role of listener familiarity in the perception of dysarthric speech. *Clinical Linguistics & Phonetics*, *9*(2), 139–154.

Weiss, M. S., & Basili, A. G. (1985). Electrolaryngeal speech produced by laryngectomized subjects: perceptual characteristics. *Journal of Speech & Hearing Research*, *28*(2), 294–300.

Weiss, M. S., Yeni-Komshian, G. H., & Heinz, J. M. (1979). Acoustical and perceptual characteristics of speech produced with an electronic artificial larynx. *The Journal of the Acoustical Society of America*, *65*(5), 1298–1308. doi:10.1121/1.382697

Whitehill, T., & Wong, C. (2006). Contributing factors to listener effort for dysarthric speech. *Journal of Medical Speech-Language Pathology*, *14*, 335–342.

Yorkston, K., & Beukelman, D. R. (1984). *Assessment of intelligibility of dysarthric speech*. Austin TX: Pro-Ed.

Zekveld, A. A., & Kramer, S. E. (2014). Cognitive processing load across a wide range of listening conditions: Insights from pupillometry. *Psychophysiology*, *51*(3), 277–284. doi:10.1111/psyp.12151

Conclusion

1. Utility of perceived listener effort as an outcome measure

This dissertation documents a path of research exploring the concept of perceived listener effort (PLE), specifically applied to disordered speech. To establish the utility of PLE as a potential outcome measure, it was necessary to show its conceptual uniqueness compared to and sensitivity to change or difference not already identified using other measures, such as speech intelligibility, overall severity of voice, and speech acceptability. As a starting point, the objective and subjective measurement of listening effort was described in Chapter 1. Considered from a hearing perspective, the factors affecting listening effort include degradation of the environment, receiver and signal source (Mattys, Davis, Bradlow & Scott, 2012). Most of the research reviewed in Chapter 1 examined effects of a degraded environment (e.g., noise level and type) or receiver (e.g., hearing impairment, age of listener). In summary, objectively measured listening effort appears to be affected by stimulus intelligibility, noise level and type, listener age, hearing status, working memory capacity, receptive vocabulary and motivation. Listening effort measured by performance is clearly related to the perception of listening effort, but there is a good deal of evidence that a reduction in performance or processing speed during challenging listening tasks is not the same as the *perception* of working very hard (Gosselin & Gagne, 2011; Zekveld et al., 2011; 2014). Moreover, ratings of PLE are only weakly correlated with objective measures of listening effort (Downs & Crum, 1978; Zekveld, Kramer & Festen, 2010). The research reviewed in this chapter strongly suggested that PLE is a different and meaningful concept from objective listening effort, and that many questions remain about the effect of a degraded signal source on PLE.

1.1 Degraded signal

The end of Chapter 1 reviewed the limited research into PLE for degraded signals (i.e., disordered or different speech). Briefly, ratings of PLE have been found to correlate strongly with transcribed intelligibility for dysarthric speech (Beukelman, Childes, Carrell, Funk, Ball & Pattee, 2011; Beukelman, Gillespie, Fager & Ullman, 2014; Whitehill & Wong, 2006), although the reported relationships were not linear. In addition to intelligibility measures, however, outcomes for disordered speech and voice are often measured using auditory-perceptual rating scales. Ratings of overall severity of voice had been shown as strongly correlated with PLE for ADSD speech (Nagle & Eadie, 2012a). The studies reported in Chapters 2 and 3 show a similarly strong, if negative, correlation between speech acceptability and PLE for alaryngeal speech. This would seem to suggest that ratings of PLE do not capture unique information about degraded or disordered speech, despite their conceptual differences from intelligibility, acceptability, etc. However, additional data from these studies indicated that there are variables that affect PLE ratings differently from acceptability or intelligibility.

1.2 Different patterns, different impressions

The potential of PLE as an outcome measure was shown using quantitative and qualitative data analysis. In Chapter 2 (Nagle & Eadie, 2012b), listeners made paired comparisons of tracheoesophageal (TE) speech, allowing ranking of the samples for acceptability and PLE. The overall rankings were very similar, but nearly every listener reversed his or her "preference" for at least one sample. That is, samples requiring more effort were not necessarily rated as less acceptable, as would be expected. These findings of individual variability in patterns of rating acceptability compared to PLE underscore the potential of PLE to provide unique information not currently measured for disordered speech. Together with the repeated reports of

a nonlinear relationship between intelligibility and PLE, findings of individual variability spurred qualitative exploration of the concepts of acceptability and PLE. Specifically, the impressions of individual inexperienced listeners seemed critical to consider when evaluating the value of PLE as an outcome measure.

The themes revealed in Chapter 4 indicate that listeners interpret acceptability and PLE as very similar, but negatively associated, concepts. Both have a strong component of "understandability" and comprehension, and a sample that is rated as requiring more effort would tend to be judged as less acceptable. However, several listeners noted a difference between these concepts. Briefly put, acceptability was how pleasant a sample was to listen to, or how tolerable a listener found it. Perceived listener effort, on the other hand, implied the burden that comes with listening to disordered speech. As intended, one participant said that "it didn't matter how pleasant a sample was to rate PLE – it only mattered how easy it was to understand what the speaker was saying". These results might suggest that at least for judging electrolaryngeal speech samples, that acceptability and PLE were related, but that these constructs can be differentiated.

1.3 Variables uniquely affecting PLE

The experiments reported in Chapter 3 showed that the sentence length of stimuli affected PLE ratings significantly more than ratings of acceptability; moreover, based on a comparison of the linear regression function and the quadratic regression function for longer sentence data, the relationship between intelligibility and PLE was nonlinear for longer (13-15 word) sentences. Based on these findings, it appears that listeners may rate acceptability as a *gestalt* regardless of its length, whereas they rate PLE as the sum of its parts. These parts seem to reflect listener (i.e. receiver) characteristics that may affect working memory and receptive

vocabulary, as described in Chapter 1. More research is needed to establish whether these characteristics affect PLE differentially.

2. Impressions of electrolaryngeal speech

Few listeners in the studies described here had ever heard electrolaryngeal (EL) speech, and only one reported ever having spoken to a person who had undergone a total laryngectomy. Participants in Chapter 4 described the speech as robotic, mechanical and lacking in prosody, all of which were expected from an electromechanical speech source. The characteristics of EL speech described by these listeners were consistent with published research on how EL speech is perceived (Bennett & Weinberg, 1973; Melztner & Hillman, 2005). However, these listeners also spoke of an affective component of listening to EL speech that may lead to the kind of social penalty described earlier this document – listeners may avoid talking to individuals using electrolaryngeal speech because they find it annoying or painful. They also found EL speech tiring and effortful to listen to. These relatively spontaneous comments about electrolaryngeal speech are consistent with quantitative findings (Evitts & Searl, 2006) when listeners process alaryngeal speech, but have not been the focus of much study to date.

3. Task differences

In the study described in Chapter 4, listeners described differences in how they rated speech samples based on the auditory-perceptual dimension (PLE or acceptability), but they also talked about using different strategies depending on the task. The task of transcribing clearly required listening for comprehension, and listeners reported working very hard to do so. They described numerous strategies such as filling in the blanks, memorizing and replaying stimuli in order to get all the words. Rating PLE was conceptually similar to transcription – comprehension was implied, although it may have taken multiple listens to decode a stimulus. Rating

acceptability of the electrolaryngeal sample, on the other hand, often meant comparing a stimulus to some internal standard of acceptable speech. In this way, listeners appeared to use the rating scale differently depending on the dimension, suggesting another way in which PLE may be different from acceptability.

This finding should be interpreted with caution, however, as the difference between transcribing and rating either task seemed to be greater than the difference between rating PLE and acceptability. Because the transcriptions were elicited in the first session only, whichever perceptual dimension was rated first may have seemed more difficult or been more focused on comprehension. The rating tasks were balanced to control an order effect, but in future studies, transcription should be done contemporaneously with rating PLE. This would provide motivation to perform the task, but more importantly, it would link the rating directly with the task being rated.

4. Future directions

The overall findings of this series of studies indicate that PLE ratings may make a unique contribution to speech and voice outcome measurement. Thus far, very little research has been done examining the effects of disordered speech on PLE. As described in Chapter 1, numerous studies have compared PLE to performance measures of listening effort for typical speech in noise; it would be prudent to examine these relationships for disordered speech as well. Likewise, the study reported in Chapter 3 is the first to simultaneously examine intelligibility, PLE and another accepted auditory-perceptual measure, acceptability. The study in Chapter 3 is also the first to report an effect of stimulus difference on PLE, but not on rated acceptability; in this case, the difference was based on length of the sentence. The effect of speech type (e.g., dysarthria vs. dysphonia vs. alaryngeal), elicitation stimulus (e.g., read vs. spontaneous speech)

and stimulus qualities other than sentence length should be explored. For example, samples used in this study, some of which were described as too fast, are currently being analyzed for speech rate (i.e., excluding pauses) to examine their effect on PLE. Effects of linguistic complexity, speaker sex and visual information could also be examined. For example, PLE might be greatly reduced with the addition of visual cues, particularly for speakers who use gestural and other nonverbal cues to enhance understanding.

Listener characteristics provide another set of obvious variables that may affect findings. The effects of age, hearing status, working memory capacity, receptive vocabulary and motivation are currently unknown for disordered speech. Given the differences between results for listening effort (performance) and PLE for hearing-impaired listeners reported by Hicks and Tharpe (2002) and for older listeners by Larsby, Hallgren, Lyxell and Arlinger (2005), exploration of these variables should be fruitful.

Another outstanding question is the appropriate scale type for measuring PLE. Inexperienced listeners were highly reliable using a paired comparison paradigm in the studies described in Chapter 2 and Nagle and Eadie (2012a); however, listeners seemed to struggle with the visual analog scale used in Chapter 3. More than half of the original listeners' data was rejected because it did not meet strict criteria for either intra-rater agreement or inter-rater reliability. It may be that an equal-appearing interval (EAI) scale provides adequate sensitivity and reliability at 9 or 11 points. The psychometric qualities of PLE are unknown; a comparison of ratings obtained via direct magnitude estimation and EAI for the same samples may shed some light on which method is best (Eadie & Doyle, 2002a; 2002b).

Finally, to obtain spontaneous speech samples and "real-world" consequences of disordered speech, future research should examine authentic conversational dyads. For a

functional outcome measure such as PLE, it is critical that the measure reflect what actually happens in a communicative interaction. Discourse analysis, for example, might reveal aspects of communicating with an individual with a speech disorder that have not arisen in controlled laboratory settings. To confirm the value of measuring PLE, it will be necessary to show that it provides meaningful information in context.

References

Bennett, S., & Weinberg, B. (1973). Acceptability ratings of normal, esophageal, and artificial larynx speech. *Journal of Speech and Hearing Research*, *16*, 608–615.

Beukelman, D., Childes, J., Carrell, T., Funk, T., Ball, L. J., & Pattee, G. L. (2011). Perceived attention allocation of listeners who transcribe the speech of speakers with amyotrophic lateral sclerosis. *Speech Communication*, *53*(6), 801–806.

Beukelman, D., Gillespie, L., Fager, S., & Ullman, C. (2014). Perceived attention allocation of listeners who transcribe the speech of dysarthric speakers with traumatic brain injury. *Journal of Medical Speech-Language Pathology*, *21*(3), 261–266.

Downs, D. W., & Crum, M. A. (1978). Processing demands during auditory learning under degraded listening conditions. *Journal of Speech and Hearing Research*, *21*(4), 702–714.

Eadie, T. L., & Doyle, P. C. (2002a). Direct magnitude estimation and interval scaling of naturalness and severity in tracheoesophageal (TE) speakers. *Journal of Speech, Language, and Hearing Research*, *45*(6), 1088–96. doi:10.1044/1092-4388(2002/087)

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

Evitts, P., & Searl, J. (2006). Reaction times of normal listeners to laryngeal, alaryngeal, and synthetic speech. *Journal of Speech, Language & Hearing Research*, *49*(6), 1380–1390.

Gosselin, P., & Gagne, J.-P. (2011). Older adults expend more listening effort than young adults recognizing speech in noise. *Journal of Speech, Language, and Hearing Research*, *54*(3), 944–958.

Hicks, C. B., & Tharpe, A. M. (2002). Listening effort and fatigue in school-age children with and without hearing loss. *Journal of Speech, Language, and Hearing Research*, *45*, 573–584.

Larsby, B., Hallgren, M., Lyxell, B., & Arlinger, S. (2005). Cognitive performance and perceived effort in speech processing tasks: Effects of different noise backgrounds in normal-hearing and hearing-impaired subjects. *International Journal of Audiology*, *44*(3), 131–143.

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes*, *27*(7-8), 953–978. doi:10.1080/01690965.2012.705006

Meltzner, G. S., & Hillman, R. E. (2005). Impact of aberrant acoustic properties on the perception of sound quality in electrolarynx speech. *Journal of Speech, Language & Hearing Research*, *48*(4), 766–79.

Nagle, K., & Eadie, T. (2012a). *Listener effort as an outcome measure for adductor spasmodic dysphonia*. Poster presented at the ASHA Convention, Atlanta, GA.

Nagle, K. F., & Eadie, T. L. (2012b). Listener effort for highly intelligible tracheoesophageal speech. *Journal of Communication Disorders*, *45*(3), 235–245. doi:10.1016/j.jcomdis.2012.01.001

Whitehill, T., & Wong, C. (2006). Contributing factors to listener effort for dysarthric speech. *Journal of Medical Speech-Language Pathology*, *14*, 335–342.

Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2011). Cognitive load during speech perception in noise: the influence of age, hearing loss, and cognition on the pupil response. *Ear and Hearing*, *32*(4), 498–510.

Zekveld, A. A., Rudner, M., Kramer, S. E., Lyzenga, J., & Rönnberg, J. (2014). Cognitive processing load during listening is reduced more by decreasing voice similarity than by increasing spatial separation between target and masker speech. *Frontiers in Neuroscience*, *8*(April), Article 88. doi:10.3389/fnins.2014.00088

Zekveld, A., Kramer, S. E., & Festen, J. M. (2010). Pupil response as an indication of effortful listening: The influence of sentence intelligibility. *Ear and Hearing*, *31*(4), 480–490.