The Effect of Speech Intelligibility on Listener Self-Perception of Effort in Electrolaryngeal Speech

Elizabeth Seagrave

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2012

Committee:

Tanya Eadie

Carolyn Baylor

Kristie Spencer

Program Authorized to Offer Degree:

Speech and Hearing Sciences

TABLE OF CONTENTS

LIST OF FIGURES

Figure Number Page

LIST OF TABLES

INTRODUCTION

Current post-laryngectomy outcomes include a variety of measures to form a comprehensive approach to assessment and treatment. Researchers in speech-language pathology view communication outcomes as a vital aspect of function for these patients. Typical outcomes include patient-reported satisfaction with speech, clinician-rated severity of voice, or judgments of speech severity by inexperienced listeners (Eadie, 2003). The focus of these outcomes remains on the function of the speaker with the communication disorder. While the performance of the speaker is critical for communication success, the reactions of the communication partner are also important to determine because of the role they play in the effectiveness of the communication dyad. A recent focus on listener factors has led to investigations providing insights into how self-perceived listener burden may play a role in speaker-patient outcomes. To date, researchers have looked at the construct of listener effort for one population of highly intelligible alaryngeal speakers (Nagle & Eadie, 2012). The purpose of this current investigation is to examine self-perceived listener effort in electrolaryngeal speech of varying intelligibility to add to the literature of listener burden in communication disorders. The following literature review will expand upon these issues.

Electrolaryngeal (EL) Speech and Factors Contributing to Optimal Intelligibility

A primary treatment for laryngeal cancer is a total laryngectomy procedure (i.e., total removal of the larynx), which leads to permanent changes in a patient's speech, swallowing, and airway function (Eadie, 2003). The three most common methods of alaryngeal speech include tracheoesophageal (TE) speech, esophageal speech, and electrolaryngeal (EL) speech. The electrolarynx, also known as an electronic artificial

larynx, is a hand-held, battery-operated device that provides an extrinsic sound source for speech production. It is commonly held on the neck (transcervical) and the sound is transmitted through the tissues of the neck, pharynx, and oral cavity. EL speech occurs when this sound source is filtered through the vocal tract, where the articulators shape the sounds for speech. The other two methods (TE and esophageal speech) are considered intrinsic, as they are methods whereby voice is produced without the use of an external device. While the driving air force is different for TE and esophageal speech, the sound source is created by the vibration of the tissues of the pharyngoesophageal (PE) segment (Doyle & Keith, 2005). TE speech employs a voice prosthesis, which allows communication between the trachea and the PE segment. It is often considered the preferred and most successful alaryngeal speech method, owing to its relative ease of acquisition and fluent, intelligible speech with relatively good voice quality (Doyle & Keith, 2005; Jongmans et al., 2010). Among experienced speech-language pathologists, TE speech was the most preferred option postlaryngectomy, while EL speech was the least preferred method (Culton & Gerwin, 1998).

Which alaryngeal speech method a patient chooses depends on a variety of factors that go beyond speech intelligibility or sound quality. The use of an electrolarynx has many benefits including its relative loudness compared to other methods of alaryngeal speech, and the fact that it can be used sooner post-surgery than other methods (Hillman, Walsh, Wolf, Fisher, & Hong, 1998). Additionally, the electrolarynx does not require as much ongoing maintenance in the way that TE speech, with its voice prosthesis, does (Farrand & Duncan, 2007). Shanks (1986) also noted that “physical factors appear to be less likely to interfere with speech using voice from a mechanical or electrical larynx” (p.

339), as compared with intrinsic methods of alaryngeal speech, thereby contributing to the electrolarynx being a "quick and easy" method to learn (Iseli et al., 2007). Indeed, Hillman et al. (1998) found that more than half of laryngectomees who use multiple methods of alaryngeal speech continue to use the electrolarynx as their primary communication method even two years postlaryngectomy.

However, a wide range of intelligibility exists in the total laryngectomy electrolarynx-using population; reported values of intelligibility range from 32% to 90% (Weiss & Basili, 1985). A variety of postlaryngectomy behavioral and physiological characteristics contribute towards an individual's ability to be understood with the electrolarynx. Slowed speech rate, overarticulation, appropriate on-off control to indicate phrasing, correct placement of the head of the device on the neck, and awareness of nonverbal communication cues are a partial list of strategies that will enhance intelligibility (Doyle & Keith, 2005). In addition, a case study by Watson and Schlauch (2009) indicated that the ability to vary fundamental frequency (f0) in EL speech (e.g., TruTone, Griffin Laboratories), in contrast to the use of more typical monotone devices, improved speech intelligibility (Watson & Schlauch, 2009). These results support the notion that use of pitch, inflection, and stress that is closest to laryngeal speech will increase the likelihood of being understood. Yet, while these factors contribute towards communication success, they only focus on the performance of the speaker as part of the communicative dyad. How these speech differences affect the listener is summarized next.

Outcomes in Alaryngeal Speech

A variety of outcome measures are used for evaluating total laryngectomy and EL speech. Traditional measurements focus on the speech signal, such as acoustic parameters and speech intelligibility (how well a speech signal is understood by others), and often are paired with auditory-perceptual judgments of speech and voice quality, along with patient-based measures examining quality of life. This multidimensional approach to evaluation is considered the standard to best evaluate outcomes after total laryngectomy (Eadie, 2007).

Research has revealed that the deviation of alaryngeal speech from listeners' internal models of typical, laryngeal speech results in perceptions of reduced speech acceptability and intelligibility (Finizia, Lindstrom, & Dotevall, 1998), owing to changes in pitch, loudness, speaking rate, stoma noise, and voice quality (Bennett & Weinberg, 1973). EL speech, specifically, has been rated by listeners as less acceptable than esophageal speech (Bennett & Weinberg, 1973). Meltzner and Hillman (2005) investigated the attributes that contribute to the unnaturalness of EL speech compared to normal natural speech and monotonous natural speech. The researchers concluded that even in conditions where known acoustic abnormalities of EL speech were improved with acoustic manipulation, certain other factors remained to negatively influence naturalness (Meltzner & Hillman, 2005), highlighting the inherent differentness of EL speech. Other researchers have found additional disadvantages with EL speech with regard to gender identification. For example, Nagle, Eadie, Wright and Sumida (2012) showed that listeners were unable to identify female EL speakers as female when they used an electrolarynx set to a low f0 (75Hz), despite an increase in intelligibility when

using the same device. The same interaction was not found for male speakers who used an electrolarynx set to a higher f0 (175Hz; i.e., listeners were always able to identify male speakers as male, regardless of the f0 setting of the device).

The social ramifications of these types of outcomes are reflected in patient-based and descriptive measures as well. A study by Danker et al. (2010) revealed that more than 85% of individuals who underwent total laryngectomies reported perceived stigmatization and social withdrawal as a result of their new voices, while more than 40% reported explicit changes to their communication behaviors, such as speaking as little as possible, leaving things unsaid, and speaking only when no other means of communication were available. As with other communication disorders, patients' feelings of ostracization from their communities and changes in their participation as a result of changes to their voices can often be the most profoundly detrimental aspect of the disorder (Eadie, 2003).

Given that communication is an interactive phenomenon and the outcomes for a population such as alaryngeal speakers depend immensely on interaction with communication partners, a truly comprehensive, multidimensional approach to outcomes in alaryngeal speech must therefore account for more than just attributes of sound quality, but also listener effects. Kreiman et al. (1993) discuss the importance of examining relationships among the signal, the task, and the listener to understand the ramifications of a listener's reaction to processing speech and how that may affect outcomes in a disordered population. These effects are especially important to consider in speakers who use EL speech, in which the signal deviates significantly from normal expectations.

Listener Effort

It is possible that the amount of effort a listener exerts while listening to a speaker is different from the listener's judgments of perceptual qualities, such as speech acceptability, naturalness, and severity (Nagle & Eadie, 2012). In fact, the variability that is seen in a listener's ratings of voice quality could be due in part to that listener's background, biases, and the interaction between these and the particular rating task (Kreiman et al., 1993). Additionally, the processing involved in decoding an incoming speech signal may be separate from the acoustic signal itself wherein a listener's perceptual judgments may be entirely different from the cognitive effort put forth to make those judgments (Evitts & Searl, 2006).

Research in communication disorders investigating listener burden is rather limited, with first investigations of "listener comfort" coming from the hearing-impaired and fluency literature (Anderson Gosselin & Gagne, 2011; O'Brian et al., 2003), and more recently in individuals with voice disorders (Eadie et al., 2007). Similarly, studies in the same literature have investigated aspects related to listener effort but focused more on the health of the hearing in the listener than in aspects of speech. Recently, Picou, Ricketts, and Hornsby (2011) explored listener effort as a function of the demand on cognitive resources when integrating multiple modalities (e.g., auditory and visual cues) during a working memory task, and found that listener effort increased in noisy conditions but was not affected by introducing visual cues.

In the area of dysarthria, researchers have examined listener effort in terms of the increased cognitive processing load placed on the listener by dysarthric speech. In a study by Whitehill and Wong (2006), listener effort and intelligibility in dysarthric Cantonese

speakers were judged by listeners. A strong correlation between the constructs (listener effort and speech intelligibility) was revealed (r_s = -.95); however, there were outliers in which equally highly intelligible samples were also judged as needing high effort, endorsing the notion that intelligibility may not encompass the same factors present in a construct of listener effort.

Similarly, a study by Beukelman, Childes, Carrell, Funk, Ball, and Pattee (2011) revealed insights into "attention allocation" of listeners who judged dysarthric speech secondary to amyotrophic lateral sclerosis. The study included speakers with mild to severe dysarthria and listeners who self-rated their attention allocation on a Likert scale as they transcribed the speakers' sentences from the *Sentence Intelligibility Test* (Yorkston et al., 2007). Results of the investigation supported the notion of a discontinuous relationship between intelligibility and perceived cognitive load, wherein attention allocation peaked between 75% and 80% intelligibility. The results provide more evidence that self-perceived listener effort taps dimensions of the communicative exchange that are not present in perceptual judgments of the signal such as intelligibility (Beukelman et al., 2011).

In a study conducted by Evitts and Searl (2006), the construct of listener effort was investigated instrumentally in application of alaryngeal speech. The researchers calculated reaction times of 60 listeners in judging matches between a single-word written stimulus and auditorily presented single-word samples of laryngeal, synthetic, and all three alaryngeal methods of speech. One highly intelligible speaker was used for each of the methods, respectively. Analysis of reaction times revealed that EL speech placed the greatest cognitive processing burden on listeners. In light of the limited scope of

stimuli (i.e., single words only), the unknown effect of using both written and auditory stimuli, and the use of only one representative sample for each method, these results should be interpreted with caution (Evitts & Searl, 2006). Future investigations should attempt to resolve and expand upon these issues.

Recently, an investigation of the perceptual, rather than instrumental/objective, construct of listener effort was introduced to potentially serve as a viable outcome measure in alaryngeal speech. Nagle and Eadie (2012) explored whether inexperienced listeners could reliably rate listener effort, and examined the uniqueness of listener effort as compared with the construct of speech acceptability in highly intelligible TE speakers. It was determined that listeners were very reliable in rating self-perceived effort and speech acceptability. Although the constructs were strongly correlated ($r > 0.99$), there were several reasons to believe that listener effort captured something that speech acceptability did not; for example, individual listener data were analyzed apart from group means and revealed that there was greater variability in the reliability of rating listener effort than there was for speech acceptability. In addition, almost all listeners judged at least one speaker as having more acceptability, but also more effort, suggesting that listeners may differ in their strategies for conceptualizing the two constructs (Nagle & Eadie, 2012).

The proposed question has obvious clinical implications for laryngectomees who use EL speech as their primary mode of communication. However, in the present study, non-laryngectomy speakers using EL speech as their sound source were purposefully chosen to test the effects of intelligibility on listener effort. Using healthy control speakers reduces any potential articulation differences or confounds on the vocal tract

that can be indicated as a result of the laryngectomy surgery (Doyle & Keith, 2005). As such, this method provides a strong initial experimental model for examining how intelligibility, as the main independent variable, could affect listener effort. Effects found with the normal, healthy system presented in this model have soundly merited implications for future studies within the total laryngectomy population.

Experimental Question

Given the results from Nagle and Eadie's (2012) study showing that listener effort is reliably rated in alaryngeal speech, an investigation is warranted to extend what we know about the construct of listener effort in a population beyond highly intelligible TE speakers. Further exploration of listener effort in EL speech is called for in light of the work by Evitts and Searl (2006), indicating that EL speech is the method that requires the most cognitive resources as measured by reaction times. Furthermore, examining a spectrum of intelligibility in this population is also worthwhile, based on results from Beukelman et al. (2011), who concluded that listener attention allocation and intelligibility of dysarthric speakers share a nonlinear relationship. Consequently, the present study was primarily designed to address the following question: How do varying levels of intelligibility affect naïve listeners' judgments of self-perceived effort in listening to EL speech?

METHODS

Participants

This study included three groups of participants: (1) 12 healthy, non-laryngectomized male speakers who provided speech samples using EL devices; (2) 30 naïve listeners who performed intelligibility testing of the speakers; and (3) 15 listeners who provided judgments of listener effort. All participants were recruited from among the Seattle, Washington community and from the undergraduate student population at the University of Washington, Seattle campus. All listeners were unfamiliar with alaryngeal speech prior to the study. Both speakers and listeners passed hearing screening tests at 25 dB at the octave frequencies between 250-4000 Hz and were proficient in English. All individuals were paid for their participation in this study. Procedures were approved by the Institutional Review Board at the University of Washington.

Speech Recordings and Stimulus Preparation

A total of 12 healthy male laryngeal speakers were selected to provide speech samples using EL devices for this study. To allow for a distributed representation across the spectrum of intelligibility and to reduce potential differences in articulation or vocal tract configuration as a result of total laryngectomy, as well as for reasons of feasibility, healthy laryngeal speakers of one gender were used. In particular, male speakers were selected because males are proportionally more represented among those who have undergone total laryngectomies (American Cancer Society, 2011). Nine speech samples were selected from among male speakers used in a previous study (Nagle et al., 2012), and three additional speakers participated in a similar protocol. Speakers' ages ranged from 20 to 69 (mean age = 28.6, SD = 13.69).

Speakers were first introduced to a monotone electrolarynx (Solatone, Griffin Laboratories) and were instructed on its use (Nagle et al., 2012). A monotone device was used for training and stimulus recording to control the effect of fundamental frequency on judgments of listener effort and intelligibility. The goal of the training was to maximize the speaker's speech output and intelligibility through proper placement and seal of the electrolarynx diaphragm against the neck, as well as to promote overarticulation. After a 10-minute training period, individuals were asked to make recordings using the electrolarynx. For the purposes of this study, only recordings from the device set at 130Hz were used. This fundamental frequency was selected because it is representative of true male EL speakers, and is close to the average f0 of laryngeal speakers (average laryngeal male f0 = 120 Hz; Doyle & Keith, 2005).

All recordings were made using a free-standing condenser microphone connected to an amplifier (Apogee Digital Trak 2) positioned twelve inches away from the speaker, with mouth-to-microphone distance held constant. Speech samples were recorded at a sampling rate of 48 kHz with 16-bit quantization, and obtained on a free-standing computer using a specialized sound card and acoustic software (Sony Soundforge 7.0). Speakers recorded six sentences of increasing length according to the *Assessment of Intelligibility of Dysarthric Speech* protocol (*AIDS;* Yorkston & Beukelman, 1984). Two sentences each that were five, six, and seven words in length were recorded. Additionally, the second sentence of the Rainbow Passage (Fairbanks, 1960) was recorded and used for listener effort ratings. Recordings were acoustically analyzed using Praat software (Boersma & Weenink, 2005) to ensure all samples had a consistent f0 of 130 Hz. Samples with an average fundamental frequency of more than 5Hz from 130Hz

were resampled using an automated Praat script to ensure that all speech samples were presented to listeners at the same average fundamental frequency. Speech samples from both the *AIDS* protocol and Rainbow Passage were also normalized for peak intensity and edited using acoustic software (Sony Soundforge 7.0). Sentences for the *AIDS* protocol were entered into the Ecos/Win software program (Avaaz Innovations, 1998), which randomly generates speaker order, presents perceptual rating scales, and records typed responses.

The 12 speech samples selected for use in this study were selected on the basis of their intelligibility, such that a broad representation across the severity continuum was observed (Hustad, 2006). Thirty listeners judged the intelligibility of samples according to the *AIDS* protocol (Yorkston & Beukelman, 1984). Listeners transcribed recognized words in the sentences, allowing for repetition of each sentence up to two times (Nagle et al., 2012). No listener heard any additional repetitions of sentences. Percent of words understood was calculated by determining the number of words correctly identified. The intelligibility score for each of the 12 speakers was based on three listeners' judgments of the six sentences from the *AIDS* protocol (controlling for sentence and word length). In summary, each speaker's intelligibility rating was based on an average of 18 judgments (3 listeners x 6 sentences per speaker); using multiple listeners decreased the effect of any outlier ratings (Shrivastav, Sapienza & Nandur, 2005). The demographic characteristics of the 12 male speakers and their average intelligibility scores are presented in Table 1.

Table 1
Speaker Demographics

Speaker #	Age	Mean Intelligibility % (SD)	f0 Hz (*prior to resampling to 130 Hz)
S1	27	99 (4.71)	130
S2	23	95 (7.14)	130
S5	21	19 (18.08)	140*
S6	20	3 (5.46)	140*
S7	20	50 (36.39)	140*
S8	24	39 (28.58)	140*
S10	28	26 (14.35)	135*
S11	24	45 (7.27)	135*
S12	28	56 (18.61)	135*
S13	38	80 (22.84)	130
S14	21	73 (9.62)	135*
S15	69	64 (11.00)	130

Listener Procedures: Judging Listener Effort

Fifteen listeners (10 females, 5 males) were recruited to participate in the second part of this study that involved judging listener effort. None of these listeners had participated in the intelligibility protocol and all were considered naïve to EL speech. The average age of the females was 24.6 years (SD = 7.75 years) and the average age of the males was 24.4 years (SD = 4.16 years).

Listeners were first familiarized with the task and given a definition of listener effort: "the amount of work needed to listen to a speaker" (Nagle & Eadie, 2012; Whitehill & Wong, 2006, p. 337). The listeners were told that they would be listening to speech produced by a speaker using an electronic device, and a brief description of the electrolarynx was provided. The samples of the second sentence of the Rainbow

Passage as described above were entered into a perceptual software program (Ruby on Rails) that presents randomized pairs of speaker stimuli to the listener and obtains listener ratings of effort using rating scales (Nagle & Eadie, 2012). Stimuli were presented to listeners over headphones (Samson Stereo Headphones, RH600) in a standard paired comparison paradigm. The paired-comparison model of stimuli presentation was selected for use in this study because it has been shown to have stronger reliability than traditional rating scales (Eadie, Doyle, Hansen & Beaudin, 2008; Meltzner & Hillman, 2005).

Listeners heard each pair once, and were asked to judge which sample of the two required more effort, using an undifferentiated 100 mm visual analog scale presented on a computer screen. In this scenario, 0mm indicates that speaker 1 requires more effort, while 100mm indicates that speaker 2 requires more effort (Appendix A). A judgment at 50mm, the middle of the line, indicates that the two samples required equal amounts of effort (Nagle & Eadie, 2012). Each listener judged all 12 speaker samples in random order of pairs two times (AxB, BxA) for listener effort, with the order of stimuli randomized across listeners (N = 12 x 11 = 132 ratings of speaker pairs). Samples within a pair were separated by 0.5 seconds (Kreiman & Gerratt, 1996; Nagle & Eadie, 2012). Listeners were able to control the rate of presentation of the pairs, as the software program required the listener to click “continue” before presenting the next pair to help control listener fatigue. However, listeners were only able to listen to the pair once for making their judgment. On average, the entire protocol took listeners 30 minutes.

Data Analysis

Raw scores for paired comparisons, measured in millimeters from the far left point (at 0 mm) of the undifferentiated 100 mm visual analog scale were converted into two ratings; one for Sample 1, and one for Sample 2. Larger numbers indicated that more listener effort was required. Accordingly, paired comparisons with values greater than 50 were interpreted as meaning Sample 2 required more effort than Sample 1. Thus, a paired comparison score of 75 would convert to a rating of "25" for Sample 1, and a rating of "75" for Speaker 2. Conversely, if a paired comparison score had a value less than 50, it was interpreted as meaning Sample 1 required more effort than Sample 2. In this case, a paired comparison score of 10 would assign a rating of "90" to Sample 1, and a rating of "10" to Sample 2 (Nagle & Eadie, 2012).

From the converted scores, an average listener effort rating for each speaker was derived (15 listeners x 132 paired comparison scores x 2 converted listener effort ratings per comparison = 3960 total listener effort ratings; 3960 divided by 12 speakers = 330 ratings of listener effort per speaker). These values were then plotted as a function of speech intelligibility scores for each speaker, and the relationship was determined using regression analyses. If visual inspection revealed a downward bowed function, then nonlinear functions were subsequently fit to the data until two non-significant results were found.

Intrarater Reliability

Twenty percent (n=13) of the speaker pairs were randomly selected to assess intrarater reliability. Pearson's correlation coefficients were determined by comparing effort ratings for AxB and BxA presentations for these speaker pairs for each listener.

Intrarater reliability values ranged from r = .48 to .93. Mean intrarater reliability was calculated to be r = .74 (SD = 0.16). Thus, intrarater reliability values were moderately high, indicating that each rater could maintain internal consistency in making listener effort judgments.

Interrater Reliability

Interrater reliability was calculated using an intraclass correlation coefficient type (3, k) (Shrout & Fleiss, 1979). The reliability for listener effort using ICC average measures was 0.957. Results indicated that reliability was adequate for ratings of listener effort for listeners as a group.

RESULTS

Speech Intelligibility and Listener Effort for Individual Speakers

Table 2 provides a summary of speech intelligibility and ratings of listener effort for each speaker. In this table, scores representing increased effort are greater (i.e., "0" = least effortful; "100" = most effortful). Because the relationship between effort and intelligibility is negative, an additional score of effort (transformed listener effort) is provided in the table such that higher scores represent less effort. The transformed scores are presented to help illustrate relationship between listener effort and intelligibility (i.e., such that for both scales, a higher score represents both "ease in listener effort" and "increased intelligibility").

Table 2
Speaker Intelligibility and Listener Effort Scores

Speaker #	Intelligibility % (SD)	Listener Effort (SD)	Transformed Listener Effort Score (100 – n)
S6	3 (5.46)	84 (8.17)	16
S5	19 (18.08)	51 (25.92)	49
S10	26 (14.35)	*58 (22.67)*	*42*
S8	39 (28.58)	*64 (19.38)*	*36*
S11	45 (7.27)	46 (28.65)	54
S7	50 (36.39)	42 (30.35)	58
S12	56 (18.61)	*51 (25.87)*	*49*
S15	64 (11.00)	47 (28.29)	53
S14	73 (9.62)	*51 (26.31)*	*49*
S13	80 (22.84)	29 (36.48)	71
S2	95 (7.14)	*39 (32.04)*	*61*
S1	99 (4.71)	36 (33.37)	64

Note: Speech intelligibility and listener effort ratings on 100mm visual analog scale, in mm, arranged from lowest to highest intelligibility score. Higher listener effort scores indicate more effort. Italics indicate reversed order for listener effort compared to intelligibility.

As shown in Table 2, listener effort ratings (most easily observed with the transformed scores) in general tended to show a strong relationship with intelligibility, with those with poor intelligibility scores coinciding with scores of increased listener effort (e.g., speaker 6), and those with high intelligibility demonstrating scores of least listener effort (e.g., speakers 1, 2, and 13). However, several individual speakers did not follow the expected order of listener effort with respect to intelligibility, with several reversals observed. For example, speakers 6 and 5 followed a trend of increasing intelligibility (3% and 19%, respectively) and corresponding decreasing listener effort scores (84 and 51, respectively), but the speaker with the next-highest intelligibility score, speaker 10 (26%), showed a reversal in listener effort rating, changing the trend (58, a listener effort score higher, rather than lower, than speaker 5's score of 51). A similar pattern is seen again between speakers 7 and 12; 15 and 14; 13 and 2. Overall, when ranked by lowest to highest intelligibility, 7 out of 12 speakers followed the expected order of decreasing listener effort, while 5 did not. Additionally of note are the speakers in the middle range of intelligibility (i.e., speakers 11, 7, 12, 15, 14; intelligibility range = 45% to 73%), who each had a mean listener effort rating that fell between 42 and 51. This nine-point range in listener effort accounted for nearly a 30% range of intelligibility. Of note are the large standard deviations for measures of listener effort within this range.

Relationship Between Speech Intelligibility and Listener Effort

Listener effort ratings plotted as a function of intelligibility scores revealed a statistically significant linear relationship ($y = -0.3978x + 71.54$, $r^2 = .656$, $F(1,10) =$

19.030, $p < .05$) (see Figure 1). These data indicate that a strong amount of the variance (65.6%) in listener effort scores is predicted by speech intelligibility ($r = .810$).

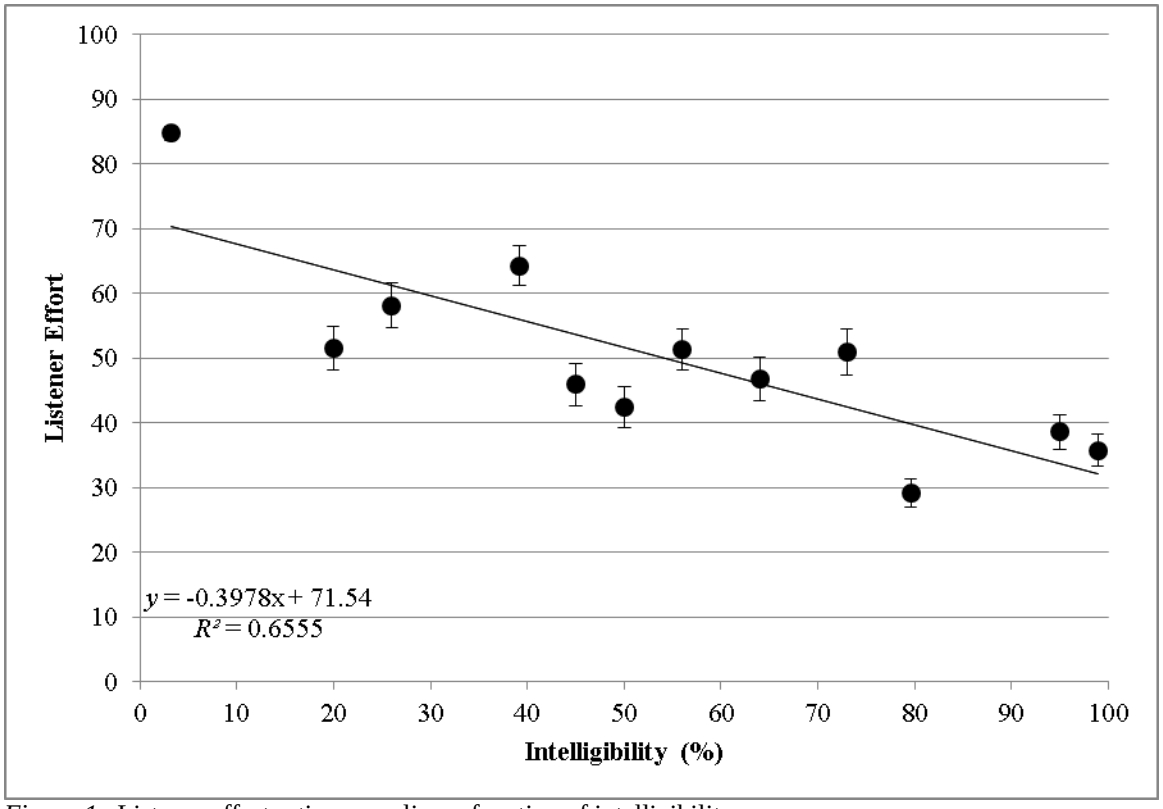

Figure 1. Listener effort ratings as a linear function of intelligibility scores.

However, visual inspection of the data appeared to reveal improvement of goodness-of-fit using a non-linear function, with the data appearing to be bowed at one end. As a result, curvilinear regression functions were subsequently fit to the data until two non-significant results were found. Results revealed that a third-order polynomial (cubic function) significantly fit the data ($y = -0.0001x^3 + 0.0224x^2 - 1.5788x + 85.885$, r^2=.745, $F(1,10)=29.243$, $p < .01$) (see Figure 2). The third-order polynomial accounted for a statistically significant amount of observed variance, above and beyond that accounted for by the simple linear and quadratic models (r^2= .745, $F(3,8)=7.798$, $p <$

.01). This result indicates that 74.5% of the variance in listener effort scores was predicted by speech intelligibility using a non-linear function (r = .863).

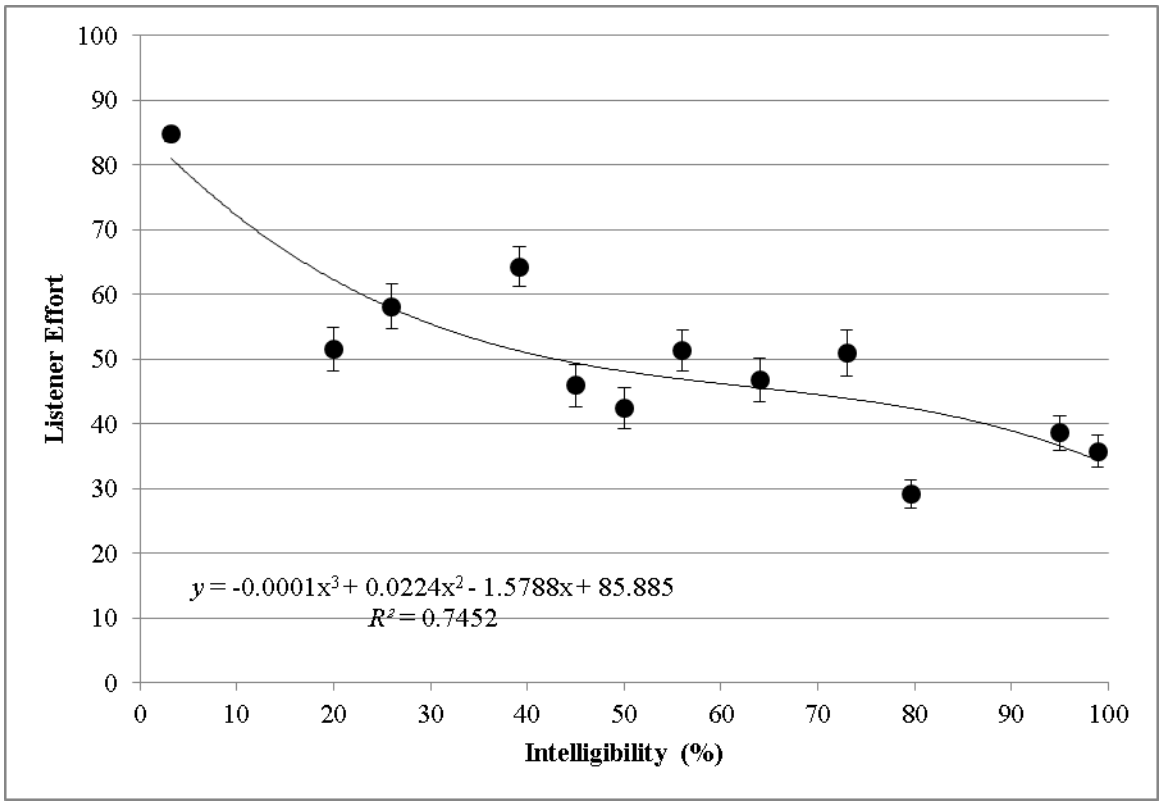

Figure 2. Listener effort ratings as a third-order polynomial function of intelligibility scores.

To further determine if there was a significant difference in listener effort between levels of intelligibility, speakers were stratified by intelligibility using the data in the best-fit cubic function (Figure 3). A one-way, between subjects ANOVA was performed to determine whether differences in listener effort were observed for speakers with different levels of intelligibility. The categorization included 4 profound speakers (intelligibility range = 3% to 39%), 5 moderate speakers (range = 45% to 73%), and 3 mild speakers (range = 80% to 99%). The results revealed a significant result ($F(2,9)=10.062$, $p < .01$) for the effect of intelligibility on listener effort. Post-hoc tests

(Tukey's) revealed significant differences for listener effort between speakers with profound severity (mean = 64.3, SD = 14.2) and those with moderate (mean = 47.4, SD = 3.8) intelligibility deficits ($p < .05$). In addition, there was a significant difference for listener effort between profound and mild speakers (mean = 34.7, SD = 5.1), indicating that speakers who were profoundly impaired for speech intelligibility were significantly more effortful to listen to than speakers in the moderate or mild categories. However, no significant difference was found for listener effort between speakers in the mild and moderate categories ($p = .172$).

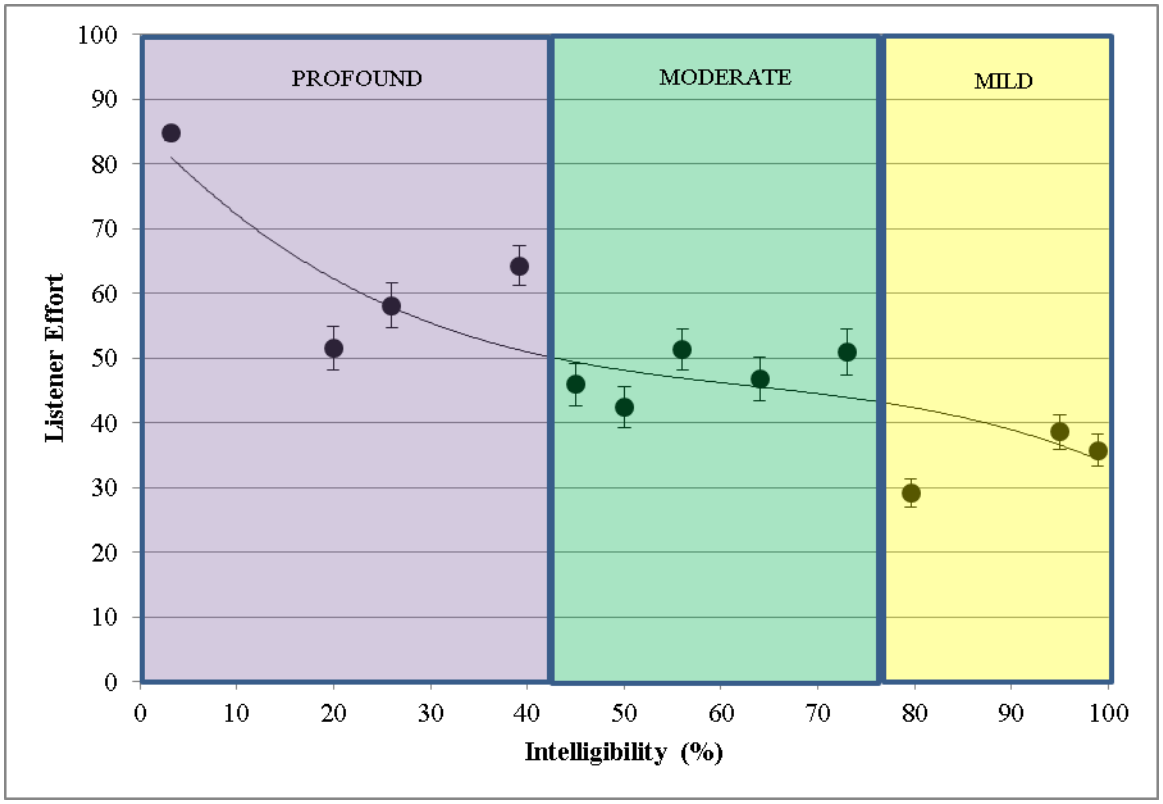

Figure 3. Scores of listener effort for each speaker stratified by intelligibility category.

DISCUSSION

The overall goal of this study was to investigate the nature of the relationship between speech intelligibility and listener effort in EL speech. To date, no research has examined the construct of listener effort specifically using this method of speech production, although there is a large body of research that has investigated listener comfort in the hearing-impaired and fluency literature (Anderson Gosselin & Gagne, 2011; O'Brian et al., 2003), and more recently in individuals with voice disorders (Eadie et al., 2007). Additionally, researchers have considered listener factors in dysarthria (Beukelman et al., 2011; Whitehill & Wong, 2006) and alaryngeal speech (Nagle & Eadie, 2012; Evitts & Searl, 2006). The studies involving speech production disorders with wide ranges of intelligibility were particularly influential in guiding the experimental question of the current study.

In general, findings of the current study were consistent with previous efforts to describe the relationship between speech intelligibility and listener effort. Specifically, while a strong correlation between speech intelligibility and listener effort was found, the relationship was best explained by a non-linear function, suggesting that listener effort is not solely a function of intelligibility for EL speech. These results will now be discussed and compared with what is known from prior literature. In particular, the significance of the best-fit curvilinear model and the implications of group comparisons will be analyzed. Following these considerations, future directions and clinical implications will be discussed, as findings from this study have the potential to influence post-laryngectomy rehabilitation, as well as methods of assessing conversational dyad outcomes in these and other communication-disordered populations.

Speech Intelligibility and Listener Effort: Individual Speaker Data

In this study, listeners used a wide range of listener effort ratings (29-84) for speakers who exhibited a wide range of speech intelligibility (3%-99%; see Table 2). The broad range of listener effort scores coupled with strong reliability for these ratings supports the contention that listeners were sensitive to differences in listener effort as a function of variability in speaker performance. Thus, similar to Nagle and Eadie (2012), the construct of listener effort appears to be a valid measure for this group of speakers who used EL devices.

There are several reasons why one can conclude that judgments of listener effort in this study were based upon factors that went above and beyond intelligibility. First, analysis of individual speaker data showed that while speakers with the highest intelligibility tended to have ratings of lowest listener effort (and vice versa), several speakers did not follow the expected order. For example, as illustrated in Table 2, a trend of increasing intelligibility with decreasing listener effort scores was seen between speakers 6 and 5, but the next speaker (10) required increased listener effort; similar patterns were observed between speakers 7 and 12; 15 and 14; 13 and 2. These results are consistent with those found by Whitehill and Wong (2006), who also reported that among their dysarthric speakers, several highly intelligible samples were also judged by some listeners as highly effortful. Thus, the present study contributes more evidence to the established notion that listeners experience a disproportionate amount of cognitive load when speech has a moderately high intelligibility. In this case, the acoustic signal may be degraded by other factors – imprecision, in the case of dysarthria, and excess

noise and additional aspects of "differentness" from laryngeal speech in the case of EL speech.

A second reason why one can conclude that listener effort is a viable construct above and beyond speech intelligibility may be found in the group data. Specifically, while a strong linear relationship was found between speech intelligibility and listener effort, the non-linear model (a cubic function) accounted for variance significantly above and beyond the linear model. This result also lends support to the conclusion that listener effort is perceptually different from intelligibility, and that listener effort potentially captures the "differentness" of alaryngeal speech (Nagle & Eadie, 2012). These results are examined next.

Speech Intelligibility and Listener Effort: Group Data

In this study, it was found that a linear model significantly fit the data (see figure 1; r = .810), indicating that a large proportion of the variance (65.6%) in listener effort ratings could be predicted by speech intelligibility scores. These results are consistent with those reported by Whitehill and Wong (2006), who also found a strong negative correlation between intelligibility and listener effort in dysarthric speech (r_s = -.95). While these data suggest that speech intelligibility strongly influences judgments of listener effort, regression analyses revealed a more complex picture. Specifically, when higher order functions were fit to the data, a third-order polynomial curvilinear function was statistically significant as a best-fit model of listener effort data plotted as a function of intelligibility (accounting for 74.5% of the variance in listener effort scores; see figure 2). This non-linear relationship accounted for variance in the data significantly above and

beyond either the linear or quadratic models. These results suggest that about 9% of the variance in scores may be attributable to something other than speech intelligibility.

Using the best-fit cubic function, speakers were further stratified by intelligibility categories (see figure 3). A between-groups comparison of listener effort data classified by deficits in intelligibility indicated statistically significant differences between profound and moderate groups, and profound and mild groups, but no difference between moderate and mild categories. In a truly linear relationship, we would expect to see differences between the mild and moderate categories with respect to listener effort. The observed lack of significance in listener effort ratings between these groups of intelligibility severity supports the notion that an increase in intelligibility does not always correspond to a decrease in listener effort, and that listeners require more effort to listen to EL speech within these moderate to mild categories of intelligibility deficits.

The results from this study are compatible with the findings from Beukelman et al. (2011), who investigated the relationship between intelligibility and listener attention allocation in dysarthric speakers with amyotrophic lateral sclerosis. These researchers found that listeners' ratings of their own attention allocation increased while intelligibility decreased from 100% to 75% (r = -.885 for speakers in this range), and peaked around 80% intelligibility. Below 75% intelligibility, attention allocation scores were mid-range. Similar to this result, the present study also revealed speakers in the middle range of intelligibility (45% to 73%) with listener effort ratings also in the mid-range (42 to 51). Likewise, when compared to other speakers in the moderate and mild categories of intelligibility, the speaker with 73% intelligibility was tied for the highest

rating of listener effort (51) across the two categories, substantiating the non-linearity of this relationship.

Results of the present study are comparable to those of Nagle and Eadie (2012), who also found that with highly intelligible TE speakers, some speakers required more listener effort than others, suggesting that aspects beyond intelligibility are contributing to these differences. Similarly, Evitts and Searl (2006) found that EL speech had greatest burden on cognitive processing for listeners compared to TE speech, esophageal speech, and synthesized speech. It is likely that the source of increased listener effort in the moderate range of intelligibility found in the current study is related to increased demands on cognitive processing owing to listeners' perception that characteristics of the acoustic signal are different from normal. These specific features of the speech (other than intelligibility), as well as listener characteristics, that affect listener processing are presently unknown and warrant further research.

While the results from this study might indicate that speakers within a particular range of intelligibility require the most amount of listener effort, there are several alternative explanations that need further study before any definitive conclusions may be made. First, a lack of significance between moderate and mild intelligibility groups might be related to the sensitivity of the perceptual scale used in capturing these differences. For example, if the scale did not include enough response options within a particular range, it would not be a valid measure. The type of perceptual scale is particularly important when measuring a perceptual dimension that is prothetic in nature. Prothetic continua are additive and quantitative in nature; a familiar example is the perception of loudness. Conversely, metathetic continua involve equal perceptual space between

intervals, such as the perception of pitch (Eadie & Doyle, 2002). Psychophysical features of voice and speech that have been found to be prothetic in nature include speech intelligibility of hearing-impaired speakers (Schiavetti, Metz, and Sitler, 1981), stuttering severity (Schiavetti et al., 1983), roughness in sustained vowels (Toner and Emanuel, 1989), nasality in synthesized vowels (Zraick and Liss, 2000) and severity of TE speech (Eadie & Doyle, 2002). Accordingly, it is possible the visual analog scale used in this study to measure listener effort did not adequately account for the nonlinearity of this dimension. However, given that the undifferentiated visual analog scale included 100 points (mm), it is unlikely that this was a factor in this study, as 1mm differences are unlikely to exceed just noticeable differences in perceiving speech attributes.

A second possible reason why relationships were found to be non-linear between speech intelligibility and listener effort could also relate to the type of scale used to measure intelligibility. For example, it is possible that the comparison between intelligibility scores in percent words correct as transcribed by listeners and ratings of listener effort is not the same as looking at a relationship between scaled intelligibility and listener effort (Sussman & Tjaden, in press). Perhaps scaled judgments of intelligibility, as well as listener effort, would be more sensitive to capturing differences and would reveal different relationships between measures. These topics should be investigated in future studies to strengthen the validity of these results.

A final reason why the present results might have revealed non-significant differences between speakers with moderate and mild speech intelligibility deficits might relate to the number of speaker samples within these ranges. For example, while there appeared to be a differential amount of listener effort used for the speaker at 73%

intelligibility in this study, it is interesting to note that the speaker with the least amount of listener effort was speaker 13, who at 80% intelligibility, was deemed to be less effortful for listeners than speaker 1 (at 99% intelligibility). This result was particularly interesting because Beukelman et al. (2011) previously found a peak for listener effort around 80% intelligibility in dysarthric speakers, although attention allocation ratings in that study were based on averages across 5 listeners who used Likert scales. In contrast, each speaker's listener effort scores were based on 330 paired comparison ratings in this study (and therefore would appear to be extremely stable). These differences in methodology might account for differences in these results across populations. The results from both the present study as well as Beukelman et al. (2011) suggest that future research should include a focused study of speakers within the moderate to mild range of intelligibility (70% to 100% intelligibility). This would allow further documentation of the relationship between intelligibility and listener effort for EL speech as well as how it might differ across speech disorders.

Future Research Directions

Future research investigations should consider the limitations of the current study. One major limitation in generalizing results from this study to those who use EL speech as their primary method of communication was the inclusion of healthy speakers who provided the speech samples. As discussed, there were several reasons why this model was selected as an initial attempt at modeling the relationship between intelligibility and listener effort. Specifically, use of this model ensured control of variables such as alterations in vocal tract effects that could affect intelligibility post-laryngectomy (Doyle & Keith, 2005). In addition, use of healthy speakers ensured an adequate breadth of

intelligibility among speakers. Although it could be argued that an intelligibility rating is an intelligibility rating across all sources, the current study would have benefitted from the ecological validity that would come from employing speakers who had undergone total laryngectomy. Future studies should examine the differences in intelligibility and effects on listener effort using members of the true disordered population.

In the present investigation, all listeners were unfamiliar with alaryngeal speech and were younger adults (average age = 24.5 years, SD = 6.6 years). While this selection criterion also permitted control of this variable in listener judgments, future studies would benefit from including listeners of all ages and experiences with this population. Recruiting older adults to provide listener effort judgments could lend ecological validity, given the fact that most caregivers and peers of laryngectomees are elderly.

Additionally, the current study examined listener effort in a controlled listening environment in which presumably all of the listener's focus was on the task. In a more realistic situation, a listener's effort may increase or decrease with other environmental factors, such as background noise or visual distractions, or other situational interferences. For example, the current study provided only auditory input to listeners, and the presence of visual input (e.g., facial expression and visual articulatory cues) could help determine the ecological validity of the results. Some researchers have noted a slight increase in intelligibility scores when visual information is added to an audio signal only, particularly for speakers with severe intelligibility deficits (Evitts, Portugal, Van Dine, & Holler, 2010; Keintz, Bunton, & Hoit, 2007). How this additional perceptual information could affect the "effort" exerted by a communication partner therefore needs future study.

The listening task used in this study included a highly structured paradigm that is often reserved for research. This type of paradigm was selected in this study as a first step to investigating relationships between speech intelligibility and listener effort so that listener reliability would be strong (Eadie et al., 2008; Meltzner & Hillman, 2005). Due to feasibility issues, paired comparisons are not often used clinically, as sample pairs and listener ratings take a significant amount of time. Therefore, results of the current study may not be directly comparable with others (e.g., Beukelman et al., 2011) that have used more traditional rating scales (e.g, seven-point Likert scales). In order for this line of research to be useful in clinical practice, relationships between measurement systems must be substantiated and warrant further investigation.

Finally, examining the sources of listener effort should be the subject of future study, and could be examined using qualitative methods. In addition, relationships between other variables above and beyond speech intelligibility (e.g., rate, intonation) that affect the speech signal should be examined to elucidate their effect on listener effort. It was not the aim of this study to qualitatively determine what listeners were taking into account when considering their effort, but anecdotally, one listener reported that it was more the "overall sound" of a sample as opposed to "how someone said a word" that made listening effortful. In addition to noting how well the speaker could be understood as a factor during the listening effort task, further comments from listeners included "background noise" as well as "the amount of speech errors" influencing ratings of effort. A systematic investigation is called for to better understand these qualities affecting perceived effort.

Clinical Implications

While this study examined the relationship between speech intelligibility and listener effort in healthy speakers who used EL devices, knowledge of this relationship has implications for EL speakers and their communication partners. It is important for EL speakers to understand that even when messages can be understood at higher levels of intelligibility, it may be the case that their communication partner's amount of effort and cognitive processing is sizeable. The current results may also explain the disinclination of a communication partner to initiate interaction or continue to interact with an EL speaker for long periods of time, or in conditions where other cognitive demands are increased. As explored by Meltzner and Hillman (2005), even when acoustically-manipulated samples of EL speech were improved with respect to known abnormalities (e.g., reduced low-frequency energy, competing noise produced by the electrolarynx itself, and increased periodicity of the sound source), listeners still perceived the speech as unnatural (Meltzner & Hillman, 2005). Current findings also help to highlight the influence of listener factors on poor social outcomes for EL speakers, such as perceived stigmatization and social withdrawal (Danker et al., 2010). It may help the EL speaker to understand that a potential negative response from a communication partner may not be due to any personal biases, but rather a function of increased load on cognitive resources. This may, in turn, help to mediate the electrolarynx user's tendency for social withdrawal or decreased participation (Eadie, 2003).

Finally, findings from this study also have implications for how we conduct rehabilitation in individuals with various types of communication disorders. As we continue to learn more about listener effort, we will take closer steps to working on

targets that improve not just speech intelligibility and other perceptual measures, but also targets to decrease listener effort. The findings of this investigation provide a rationale for and contribution towards an evidence base for understanding how listener effort contributes towards outcomes. The results also provide the impetus for improving awareness and lessening the burden on one's communication partner, such as incorporating alternative methods and strategies to supplement verbal communication. Ultimately, it is hoped that this will lead to better care for those with communication disorders.

REFERENCES

American Cancer Society. (2001). What are the risk factors for laryngeal and hypopharyngeal cancers? Retrieved from http://www.cancer.org/Cancer/LaryngealandHypopharyngealCancer/DetailedGuide/laryngeal-and-hypopharyngeal-cancer-risk-factors.

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.

Avaaz Innovations Inc. (1998). Experiment Controller and Generator for Windows (Ecos/Win) [Computer software]. London, Ontario, Canada: Author.

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.

Boersma, P., & Weenink, D. (2005). Praat: Doing Phonetics by Computer (Version 4.3.01) [Computer software]. Retrieved from http://www.praat.org/

Culton, G. L., & Gerwin, J. M. (1998). Current trends in laryngectomy rehabilitation: A survey of speech-language pathologists. *Otolaryngology – Head and Neck Surgery, 118*, 458-463.

Danker, H., Wollbruck, D., Singer, S., Fuchs, M., Brahler, E., & Meyer, A. (2010). Social withdrawal after laryngectomy. *European Archives of Otorhinolaryngology, 267*(4), 593-600.

Doyle, P. C., & Keith, R. L. (Eds.). (2005). *Contemporary Considerations in the Treatment and Rehabilitation of Head and Neck Cancer*. Austin, TX: Pro-Ed.

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*(6), 3014-3021.

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.

Eadie, T. L., Nicolici, C., Baylor, C. R., Almand, K., Waugh, P., & Maronian, N. (2007). Effect of experience on judgments of adductor spasmodic dysphonia (ADSD). *Annals of Otology, Rhinology and Laryngology*, *116*(9), 695-701.

Evitts, P. M., Portugal, L., Van Dine, A., & Holler, A. (2010). Effects of audio-visual information on the intelligibility of alaryngeal speech. *Journal of Communication Disorders, 43*(2): 92-104.

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.

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

Farrand, P., & Duncan, F. (2007). Generic health-related quality of life amongst patients employing different voice restoration methods following total laryngectomy. *Psychology, Health & Medicine*, *12*(3), 255-265.

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

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.

Iseli, T. A., Agar, N. M., Dunemann, C., & Lyons, B. M. (2007). Functional outcomes following total laryngopharygectomy. *ANZ Journal of Surgery*, *77*(11), 954-957.

Jongmans, P., Wempe, T., Tinteren, H., Hilgers, F. J., Pols, L. C., et al. (2010). Acoustic analysis of the voiced-voiceless distinction in dutch tracheoesophageal speech. *Journal of Speech, Language, and Hearing Research, 53*: 284-297.

Keintz, C. K., Bunton, K., & Hoit, J. D. (2007). Influence of visual information on the intelligibility of dysarthric speech. *American Journal of Speech-Language Pathology, 16*(3), 222-234.

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*(1), 21-40.

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.

Nagle, K. F., & Eadie, T. L. (2012). Listener Effort for Highly Intelligible Tracheoesophageal Speech. *Journal of Communication Disorders, 45,* 235-245.

Nagle, K. F., Eadie, T. L., Wright, D. R., and Sumida, Y. A. (2012). Effect of fundamental frequency on judgments of electrolaryngeal speech. *American Journal of Speech-Language Pathology, 24,* 154-166.

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.

Picou, E. M., Ricketts, T. A., & Hornsby, B. W. Y. (2011). Visual cues and listening effort: Individual variability. *Journal of Speech, Language, and Hearing Research, 54,* 1416-1430.

Schiavetti, N., Metz, D. E., and Sitler, R. W. (1981). Construct validity of the direct magnitude estimation and interval scaling: Evidence from a study of the hearing-impaired. *Journal of Speech and Hearing Research, 24,* 441-445.

Schiavetti, N., Sacco, P. R., Metz, D. E., and Sitler, R. W. (1983). Direct magnitude estimation and equal appearing interval scaling of stuttering severity. *Journal of Speech and Hearing Research, 26,* 568-573.

Shanks, J. C. (1986). Essentials for alaryngeal speech: Psychology and physiology. In R. L. Keith & F. L. Darley (Eds.), *Laryngectomee Rehabilitation* (337-349). Austin: Pro-Ed.

Shrivastav, R., Sapienza, C. M., & Nandur, V. (2005). Application of psychometric theory to the measurement of voice quality using rating scales. *Journal of Speech, Language, and Hearing Research*, *48*(2), 323-35.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychology Bulletin, 86,* 420–428.

Sussman, J. E., & Tjaden, K. (in press). Perceptual measures of speech from individuals with Parkinson's Disease and Multiple Sclerosis: Intelligibility and beyond. *Journal of Speech, Language, and Hearing Research.* Retrieved from doi:10.1044/1092-4388(2011/11-0048

Toner, M. A., and Emanuel, F. W. (1989). Direct magnitude estimation and equal appearing interval scaling of vowel roughness. *Journal of Speech and Hearing Research, 32,* 78-82.

Watson, P. J., & Schlauch, R. S. (2009). Fundamental frequency variation with an electrolarynx improves speech understanding: A case study. *American Journal of Speech-Language Pathology, 18*(2), 162–167.

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

Whitehill., T. L., & Wong, C. C-Y. (2006). Contributing factors to listener effort for dysarthric speech. *Journal of Medical Speech-Language Pathology, 14*(4), 335-341.

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

Yorkston, K., Beukelman, D., Hakel, M., & Dorsey, M. (2007). Sentence Intelligibility Test, Institute for Rehabilitation Science and Engineering at Madonna Rehabilitation Hospital, Lincoln NE.

Zraick, R. I., and Liss, J. M. (2000). A comparison of equal-appearing interval scaling and direct magnitude estimation of nasal voice quality. *Journal of Speech, Language, and Hearing Research, 43,* 979-988.

APPENDIX A

Instructions for listener effort task:

You will be listening to speech samples from adult males. The speech is produced with an electrolarynx, which is an electronic battery-operated device that provides a sound source for the speaker.

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 MORE effort for you to listen to, and by how much. For example, if you perceive Speaker #1's voice to require MORE 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 MORE 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 MORE effort than the other, move the cursor farther toward that end.