Published in final edited form as:
J Commun Disord. 2012 May ; 45(3): 235–245. doi:10.1016/j.jcomdis.2012.01.001.

Listener Effort for Highly Intelligible Tracheoesophageal Speech

Kathy F. Nagle and
Department of Speech & Hearing Sciences, University of Washington

Tanya L. Eadie
Department of Speech & Hearing Sciences, University of Washington

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

Correspondence concerning this article should be addressed to Kathy Nagle, Department of Speech and Hearing Sciences, University of Washington, 1417 NE 42nd St. Seattle, WA 98105. kfnagle@uw.edu; Phone: (206) 685-7615; Fax (206) 543-1093.

Note: Portions of this paper were presented at the *Annual Convention of the American Speech-Language-Hearing Association*, November 2010, Philadelphia, PA.

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" 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 listener effort 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 listener effort 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

listener effort for some speakers supports the idea that "listener effort" 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 listener effort 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 listener effort 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 listener effort. 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 listener effort 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 listener effort in TE speech?
2. Is listener effort a viable construct in TE speech? That is, does 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 listener effort 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 listener effort and speech acceptability to ensure the selected samples displayed a range of each dimension. To control possible effects of dialect on intelligibility, acceptability or listener effort, 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 ($n = 14 \times 13 = 182$) 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 listener effort (see Appendix). For the purposes of this study, listener effort 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 (listener effort 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 Fthe other sample in the pair. Likewise, a sample with a converted score of 25 was interpreted as less acceptable or requiring more listener effort 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 × 2 stimulus orders × 20 listeners = 520 judgments per speaker per dimension).

To answer the experimental question of whether inexperienced listeners can reliably judge listener effort, 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 listener effort, 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 listener effort and for speech acceptability. The relationship was determined using a Pearson's Correlation Coefficient. Matched pair *t* tests with Bonferonni corrections ($p < .0025$) were also calculated for individual listener data to determine the significance of individual listener differences in ratings of listener effort 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 listener effort, 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 listener effort 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 listener effort to speech acceptability. Using the rating scales provided, a lower score indicated more listener effort, 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 listener effort and from 20.91–80.90 for speech acceptability, as shown in Table 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).

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 listener effort, 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, there were some large differences in reliability between the dimensions for some listeners, although the difference between group average reliability for acceptability and listener effort was not significant [$t(19) = −0.15$; $p = .883$].

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 listener effort ($r = .66$) 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 listener effort ($r = .98$) and for speech acceptability ($r = .98$).

The sample variance for listener effort 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 1.

3.4 Relationship between Listener Effort and Speech Acceptability

The relationship between the two dimensions was determined to establish whether ratings of listener effort 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 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).

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 listener effort 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 listener effort 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 listener effort.

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).

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 listener effort.

Notably, a range of mean discrete ratings was assigned to both acceptability (20.91–80.90) and listener effort (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 listener effort 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 listener effort 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 listener effort for some listeners. Though not significant, individual listeners tended to rate speech acceptability of TE speakers more reliably than listener effort. These results suggest that the term or the concept of acceptability might have more perceptual reality for inexperienced listeners than a construct such as "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 listener effort 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 listener effort ($r = .50$; see Table 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

listener effort (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 listener effort 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 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 listener effort. 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 "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 listener effort add limited information to currently used outcome measures. However, to determine whether ratings of listener effort add any independent information, it is first necessary to examine the relationship between these measures.

4.2 Relationship between Listener Effort and Speech Acceptability

To determine whether ratings of listener effort 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 listener effort 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 listener effort 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 listener effort 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 listener effort 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 listener effort 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 listener effort 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 listener effort 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 listener effort; 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 listener effort 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 listener effort, 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 listener effort (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 listener effort 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 listener effort, 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 listener effort 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 listener effort may guide clinicians in choosing a focus of treatment; for example, a treatment method that simultaneously increases intelligibility and reduces listener effort 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 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 listener effort and investigate its relationship to other auditory-perceptual dimensions of speech and voice.

References

American Cancer Society. Cancer Facts and Figures. 2010. Retrieved from http://www.cancer.org/acs/groups/content/@nho/documents/document/acspc-024113.pdf

ANSI. ANSI S1.1–1960, Acoustical terminology. American National Standards Institute; New York: 1960.

Anderson Gosselin P, Gagné J-P. Older adults expend more listening effort than young adults recognizing speech in noise. Journal of Speech, Language, and Hearing Research. 2011; 54:944–958.10.1044/1092–4388(2010/10–0069)

Bennett S, Weinberg B. Acceptability ratings of normal, esophageal, and artificial larynx speech. Journal of Speech and Hearing Research. 1973; 16:608–615. [PubMed: 4783798]

Beukelman DR, Childes J, Carrell T, Funk T, Ball LJ, Pattee GL. Perceived attention allocation of listeners who transcribe the speech of speakers with amyotrophic lateral sclerosis. Speech Communication. 2011; 53:801–806.10.1016/j.specom.2010.12.005

Butler, EH. Acoustic analysis of voice quality: A tabulation of algorithms 1902–1990. In: Kent, RD.; Ball, MJ., editors. Voice Quality Measurement. San Diego: Singular Publishing Group; 2000. p. 119-244.

Chan KM, Yiu EM. The effect of anchors and training on the reliability of perceptual voice evaluation. Journal of Speech, Language, and Hearing Research. 2002; 45:111–126.10.1044/1092–4388(2002/009)

DeBodt MS, Wuyts FL, VandeHeyning PH, Croux C. Test-retest study of the GRBAS scale: Influence of experience and professional background on perceptual rating of voice quality. Journal of Voice. 1997; 11:74–80. [PubMed: 9075179]

Derwing TM, Munro MJ. Second language accent and pronunciation teaching: A research-based approach. Tesol Quarterly. 2005; 39:379–397.

Eadie TL. The ICF: a proposed framework for comprehensive rehabilitation of individuals who use alaryngeal speech. American Journal of Speech-Language Pathology. 2003; 12:189–197. [PubMed: 12828532]

Eadie TL. Application of the ICF in communication after total laryngectomy. Seminars in Speech and Language. 2007; 28:291–300. [PubMed: 17935014]

Eadie TL, Doyle PC. Direct magnitude estimation and interval scaling of pleasantness and severity in dysphonic and normal speakers. Journal of the Acoustical Society of America. 2002; 112:3014–3021. [PubMed: 12509023]

Eadie TL, Doyle PC. Auditory-perceptual scaling and quality of life in tracheoesophageal speakers. Laryngoscope. 2004; 114:753–759. [PubMed: 15064636]

Eadie TL, Doyle PC. Quality of life in male tracheoesophageal (TE) speakers. Journal of Rehabilitation Research and Development. 2005; 42:115–124. [PubMed: 15742256]

Eadie TL, Doyle PC, Hansen K, Beaudin PG. Influence of speaker gender on listener judgments of tracheoesophageal speech. Journal of Voice. 2008; 22:43–57. S0892–1997(06)00111–1 [pii]. 10.1016/j.jvoice.2006.08.008 [PubMed: 17055223]

Eadie TL, Nicolici C, Baylor C, Almand K, Waugh P, Maronian N. Effect of experience on judgments of adductor spasmodic dysphonia. Annals of Otology, Rhinology and Laryngology. 2007; 116:695–701.

Evans D, Healey EC, Kawai N, Rowland S. Middle school students' perceptions of a peer who stutters. Journal of Fluency Disorders. 2008; 33:203–219. S0094–730X(08)00034-X [pii]. 10.1016/j.jfludis.2008.06.002 [PubMed: 18762062]

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

Fairbanks, G. Voice and articulation drillbook. 2. New York: Harper; 1960.

Finizia C, Lindström J, Dotevall H. Intelligibility and perceptual ratings after treatment for laryngeal cancer: laryngectomy versus radiotherapy. Laryngoscope. 1998; 108:138–143. [PubMed: 9432084]

Gerratt BR, Kreiman J. Measuring vocal quality with speech synthesis. Journal of the Acoustical Society of America. 2001; 110:2560–2566. [PubMed: 11757945]

Gerratt BR, Kreiman J, Antonanzas-Barroso N, Berke GS. Comparing internal and external standards in voice quality judgments. Journal of Speech and Hearing Research. 1993; 36:14–20. [PubMed: 8450655]

Healey EC, Gabel RM, Daniels DE, Kawai N. The effects of self-disclosure and non self-disclosure of stuttering on listeners' perceptions of a person who stutters. Journal of Fluency Disorders. 2007; 32:51–69. S0094–730X(07)00002–2, [pii]. 10.1016/j.jfludis.2006.12.003 [PubMed: 17275902]

Iversen-Thoburn SK, Hayden PA. Alaryngeal speech utilization: A survey. Journal of Medical Speech-Language Pathology. 2000; 8:85–99.

Klasner ER, Yorkston KM. Speech intelligibility in ALS and HD dysarthria: The everyday listener's perspective. Journal of Medical Speech-Language Pathology. 2005; 13:127–139.

Kreiman J, Gerratt BR. The perceptual structure of pathologic voice quality. Journal of the Acoustical Society of America. 1996; 100:1787–1795. [PubMed: 8817904]

Kreiman J, Gerratt BR. Validity of rating scale measures of voice quality. Journal of the Acoustical Society of America. 1998; 104:1598–1608. [PubMed: 9745743]

Kreiman J, Gerratt BR. Perception of aperiodicity in pathological voice. Journal of the Acoustical Society of America. 2005; 117:2201–2211. [PubMed: 15898661]

Kreiman J, Gerratt BR, Kempster GB, Erman A, Berke GS. Perceptual evaluation of voice quality: review, tutorial, and a framework for future research. Journal of Speech and Hearing Research. 1993; 36:21–40. [PubMed: 8450660]

Kreiman J, Gerratt BR, Precoda K. Listener experience and perception of voice quality. Journal of Speech and Hearing Research. 1990; 33:103–115. [PubMed: 2314068]

Kreiman J, Gerratt BR, Precoda K, Berke GS. Individual differences in voice quality perception. Journal of Speech and Hearing Research. 1992; 35:512–520. [PubMed: 1608242]

Kreiman, J.; Gerratt, B. Measuring vocal quality. In: Kent, RD.; Ball, MJ., editors. Voice Quality Measurement. San Diego: Singular Publishing Group; 2000. p. 73-101.

National Cancer Institute. Laryngeal Cancer Treatment. 2010. Retrieved from http://www.cancer.gov/cancertopics/pdq/treatment/laryngeal/Patient/page4

Maccallum JK, Cai L, Zhou L, Zhang Y, Jiang JJ. Acoustic analysis of aperiodic voice: perturbation and nonlinear dynamic properties in esophageal phonation. Journal of Voice. 2009; 23:283–290. S0892–1997(07)00133–6, [pii]. 10.1016/j.jvoice.2007.10.004 [PubMed: 18411036]

Maryn Y, Dick C, Vandenbruaene C, Vauterin T, Jacobs T. Spectral, cepstral, and multivariate exploration of tracheoesophageal voice quality in continuous speech and sustained vowels. Laryngoscope. 2009; 119:2384–2394.10.1002/lary.20620 [PubMed: 19718753]

McDonald R, Parsa V, Doyle PC. Objective estimation of tracheoesophageal speech ratings using an auditory model. Journal of the Acoustical Society of America. 127:1032–1041.10.1121/1.3270396

Meltzner GS, Hillman RE. Impact of aberrant acoustic properties on the perception of sound quality in electrolarynx speech. Journal of Speech, Language and Hearing Research. 2005; 48:766–779.10.1044/1092–4388(2005/053)

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

Pindzola RH, Cain BH. Acceptability ratings of tracheoesophageal speech. Laryngoscope. 1988; 98:394–397. [PubMed: 3352438]

Rakerd B, Seitz FP, Whearty M. Assessing the cognitive demands of speech listening for people with hearing losses. Ear & Hearing. 1996; 17:97–106. Retrieved from http://journals.lww.com/ear-hearing/pages/default.aspx. [PubMed: 8698163]

Robbins J, Fisher HB, Blom EC, Singer MI. A comparative acoustic study of normal, esophageal, and tracheoesophageal speech production. Journal of Speech and Hearing Disorders. 1984; 49:202–210. [PubMed: 6716991]

Searl JP, Small LH. Gender and masculinity-femininity ratings of tracheoesophageal speech. Journal of Communication Disorders. 2002; 35:407–420. [PubMed: 12194562]

Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin. 1979; 86(2):420–8. [PubMed: 18839484]

Shrivastav R, Sapienza CM, Nandur V. Application of psychometric theory to the measurement of voice quality using rating scales. Journal of Speech, Language, and Hearing Research. 2005; 48:323–335.

Singer MI, Blom ED. An endoscopic technique for restoration of voice after laryngectomy. Annals of Otology, Rhinology and Laryngology. 1980; 89:529–533.

Trudeau MD. A comparison of the speech acceptability of good and excellent esophageal and tracheoesophageal speakers. Journal of Communication Disorders. 1987; 20:41–49. [PubMed: 3819002]

Trudeau MD, Qi YY. Acoustic characteristics of female tracheoesophageal speech. Journal of Speech and Hearing Disorders. 1990; 55:244–250. [PubMed: 2329786]

van As-Brooks CJ, Koopmans-van Beinum FJ, Pols LC, Hilgers FJ. Acoustic signal typing for evaluation of voice quality in tracheoesophageal speech. Journal of Voice. 2006; 20:355–368. S0892–1997(05)00060–3, [pii]. 10.1016/j.jvoice.2005.04.008 [PubMed: 16185840]

van As CJ, Hilgers FJ, Verdonck-de Leeuw IM, Koopmans-van Beinum FJ. Accoustical Analysis and perceptual evaluation of tracheoesophageal prosthetic voice. Journal of Voice. 1998; 12:239–248. [PubMed: 9649080]

Whitehill TL, Wong CC-Y. Contributing factors to listener effort for dysarthric speech. Journal of Medical Speech-Language Pathology. 2006; 14:335–341.

Yorkston KM, Strand EA, Kennedy MRT. Comprehensibility of dysarthric speech: implications for assessment and treatment planning. American Journal of Speech-Language Pathology. 1996; 5:55–66.

Zekveld AA, Kramer SE, Festen JM. Pupil response as an indication of effortful listening: The influence of sentence intelligibility. Ear & Hearing. 2010; 31:480–490.10.1097/AUD.0b013e3181d4f251 [PubMed: 20588118]

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.

HIGHLIGHTS

- Inexperienced listeners reliably judged listener effort in tracheoesophageal speech.
- Ratings of listener effort and of speech acceptability were highly correlated.
- Equally intelligible samples received a range of ratings for effort and acceptability.
- Listener effort may capture information not expressed in ratings of accetpability.

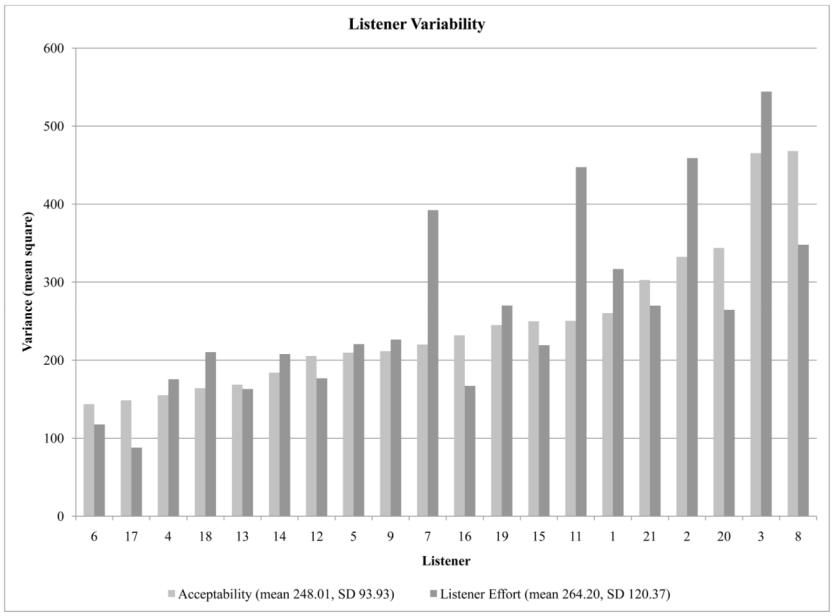

Figure 1.
Interrater variability for the dimensions of speech acceptability and listener effort, expressed in mean squares for each listener.

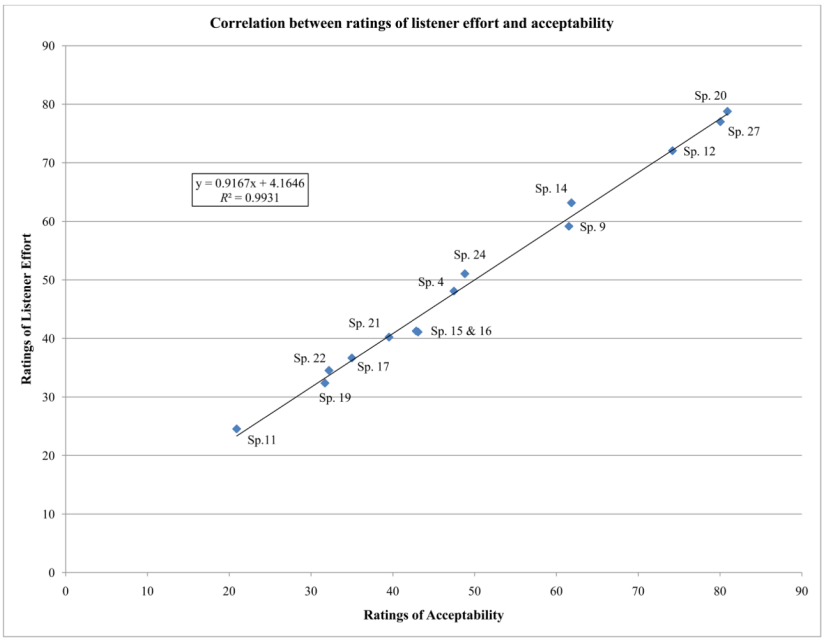

Figure 2.
Correlation between mean discrete scores in mm (0–100) for listener effort and acceptability, by speaker.

Table 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

Speaker #	Acceptability Mean (SD)	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)

Table 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)