Direct Magnitude Estimation and Interval Scaling of Naturalness and Severity in Tracheoesophageal (TE) Speakers

Tanya L. Eadie
Philip C. Doyle
School of Communication Sciences and Disorders
Elborn College
University of Western Ontario
London, Ontario, Canada

The purpose of this study was to determine the psychophysical character and validity of auditory-perceptual ratings of naturalness and overall severity for tracheoesophageal (TE) speech. This was achieved through use of direct magnitude estimation (DME) and equal-appearing interval (EAI) scaling procedures. Twenty adult listeners judged speech naturalness and overall severity from connected speech samples produced by 20 adult male TE speakers. A comparison of DME- and EAI-scaled judgments yielded a metathetic continuum for naturalness and a prothetic continuum for overall severity. These data provide support for the use of either DME or EAI scales in auditory-perceptual ratings of naturalness, but they provide support only for DME scales in judging overall severity for TE speech. The present results suggest that the nature of perceptual phenomena (prothetic vs. metathetic) for TE speakers is consistent with findings for the same dimensions produced by normal laryngeal speakers. These data also support a need for further study of perceptual dimensions associated with TE voice and speech in order to avoid the inappropriate and invalid use of EAI scales frequently found in diagnosis, assessment, and evaluation of this clinical population.

KEY WORDS: tracheoesophageal speech, perceptual scaling, alaryngeal speech, naturalness, severity

Despite a long history in the psychophysical literature (cf. Stevens, 1975) for the application of direct magnitude estimation (DME), the DME method is re-emerging in relation to perceptual judgments of various parameters of voice and speech. DME scales permit listeners to make perceptual judgments in relation to a "standard" that is designed to represent the approximate midpoint of any given set of stimuli in a perceptual continuum (Schiavetti, 1984). This standard is termed a "modulus" and is usually given a value of 100. For example, in an auditory-perceptual task, listeners are asked to judge the voice or speech of a group of speakers (e.g., individuals with dysphonia or dysarthria) for a given perceptual attribute (e.g., roughness or intelligibility). Then speech samples are judged (i.e., a numerical rating value is given) for the specified perceptual attribute. Individual samples with a voice or speech attribute that is judged to be twice as good as the modulus are assigned a value of 200, whereas speaker samples judged to have a voice or speech attribute only half as good as the modulus are assigned a

value of 50. As such, the endpoints of the DME continua are unspecified, as perceptual phenomena are scaled relative to the modulus.

In contrast, equal-appearing interval (EAI) scales require listeners to provide perceptual ratings based on a fixed, predefined scale that suggests implied "equality" of perceptual distance, weight, or magnitude between numeric components. In the voice literature, most EAI scales are 7 points (i.e., 7 pt-EAI), with "1" (e.g., normal voice quality) representing one end of the scale and "7" (e.g., severely impaired voice quality) representing the other extreme (Kreiman, Gerratt, Kempster, Erman, & Berke, 1993). Other common EAI scales used in the literature include 5-point and 9-point EAI scales (Kreiman et al., 1993). Unlike DME, the endpoints of EAI scales are fixed, and scaling is performed using whole number representatives (i.e., any whole number between "1" and "n").

Stevens (1975) has indicated that there are basically two types of perceptual continua that can be scaled: *prothetic* and *metathetic*. A prothetic continuum is additive and *quantitative* in nature. It is best scaled using DME because observers cannot subdivide a prothetic continuum into equal intervals. In contrast, a metathetic continuum is a substitutive, *qualitative* continuum that can be scaled using either DME or EAI scaling procedures. The prototypical example for a prothetic continuum is loudness, whereas pitch best exemplifies a metathetic continuum. Stevens (1975) outlined a method for determining whether a given dimension falls along a metathetic or prothetic continuum. In this procedure, the arithmetic means of the EAI ratings of a scale are plotted against the geometric means of the DME scores. If the relationship between these means is linear, then the scale is considered metathetic in nature, implying equal perceptual space between the intervals of the scale. Metathetic dimensions, therefore, may be scaled using either DME or EAI scales. However, if the relationship between the EAI scores and DME scores is nonlinear, it is suggestive of a prothetic continuum, for which only the DME method is appropriate.

Researchers in communication sciences and disorders have tested a variety of psychophysical attributes of voice and speech using Stevens' (1975) procedure. Perceptual dimensions commonly scaled in voice and speech often have revealed prothetic continua, suggesting that for these dimensions, DME rating scales are most appropriate. Using the methodology of Stevens (1975) and others (e.g., Barry & Kidd, 1981), the speech intelligibility of speakers with hearing loss (Schiavetti, Metz, & Sitler, 1981), stuttering severity (Schiavetti, Sacco, Metz, & Sitler, 1983), judgments of roughness for sustained vowels (Toner & Emanuel, 1989), ratings of nasality for synthesized vowels (Zraick & Liss, 2000), and ratings of hypernasality in connected speech samples of individuals with repaired cleft palate (Whitehill, Lee, & Chun, 2002) have been found to be prothetic. However, in examining the acoustic and psychophysical dimensions of perceived speech naturalness of nonstutterers and posttreatment stutterers, Metz, Schiavetti, and Sacco (1990) found that speech naturalness behaves like a metathetic continuum. Sewall, Weglarski, Metz, Schiavetti, and Whitehead (1999) found that the ratings of breathiness in normal speakers also are metathetic. Because EAI scales abound in the speech-language pathology literature (cf. Kreiman et al., 1993), it is critical that the nature (i.e., prothetic or metathetic) of perceptual dimensions is determined such that appropriate and valid scales are used. The validity of such auditory-perceptual scales has widespread implications for diagnosis and treatment outcomes, as perceptual scales are used most often and valued most highly by clinicians (Gerratt, Till, Rosenbek, Wertz, & Boysen, 1991).

To date, the construct validity of DME versus EAI scaling has only been investigated for dimensions of speech and voice in laryngeal speakers, with only limited work undertaken that concerns the voice/speech of the postlaryngectomy "alaryngeal" speaker. However, due to the unique nature of alaryngeal speech, the relevance of Stevens' (1975) procedures continues to raise questions concerning listener judgments of inherent perceptual dimensions. Such work has the potential to serve as a frame of reference for determining one or more aspect(s) of postlaryngectomy "outcomes" for individuals who use alaryngeal speech (i.e., esophageal, tracheoesophageal, or artificial laryngeal methods of postlaryngectomy communication).[1]

Since its development in 1980, the tracheoesophageal (TE) puncture technique has offered an option for postlaryngectomy voice and speech production (Singer & Blom, 1980). In the TE puncture procedure, a fistula is surgically created so that there is a connection between the trachea and the esophagus (which serves as the vicarious voice source). When air is exhaled and the tracheostoma is occluded, pulmonary air is shunted through a TE voice prosthesis into the esophageal reservoir, where it then creates oscillation of the pharyngoesophageal (PE) sphincter. Oscillation of the PE sphincter then creates the alaryngeal voice source that is used for speech production (Blom, Singer, & Hamaker, 1986). A clear advantage of the TE puncture approach is that there is access to pulmonary air for voice production, thereby permitting higher trans-pseudoglottal airflow values (Moon & Weinberg, 1987). Higher airflow

[1] It is beyond the scope of the present paper to describe each postlaryngectomy voice and speech option. For a more thorough presentation of this information, the reader is referred to texts by Blom, Singer, and Hamaker (1998); Doyle (1994); Keith and Darley (1986, 1994); and Snidecor (1978).

rates subsequently affect pitch (i.e., higher than traditional esophageal speech), with some researchers finding a nonsignificant difference between normal laryngeal speakers and TE speakers for the acoustic measure of fundamental frequency (Finizia, Dotevall, Lundström, & Lindström, 1999; Hillman, Walsh, Wolf, Fisher, & Hong, 1998; Robbins, Fisher, Blom, & Singer, 1984; Trudeau & Qi, 1990; Van As, Hilgers, Verdonck-de Leeuw, & Koopmans-van Beinum, 1998). Access to pulmonary air also permits a speech rate that is comparable to normal laryngeal speakers (Pindzola & Cain, 1989). When compared to the other methods of alaryngeal voice and speech production, TE speakers are usually judged as being most "intelligible and pleasant" and are among those that exhibit frequency, intensity, and durational values closest to normal when compared with other alaryngeal methods (cf. Hillman et al., 1998; Pindzola & Cain, 1988, 1989; Robbins et al., 1984; Van As et al., 1998; Williams & Watson, 1985; and others).

Although speech acceptability has been shown to be high in TE speakers, it must be emphasized that alaryngeal methods may be judged to be less acceptable and TE speakers have been rated as exhibiting decreased voice quality relative to normal laryngeal speakers, as well as those speakers who have undergone radiation therapy (Finizia et al., 1999; Tardy-Mitzell, Andrews, & Bowman, 1985). These studies illustrate the following principle—although alaryngeal modes of communication have improved in the past few decades, alaryngeal speakers can still be identified as "different" from normal speakers, and variability between speakers within any given alaryngeal group (e.g., TE speakers) will exist. Hence, it may be postulated that more holistic auditory-perceptual judgments of alaryngeal speech in general, and TE speech in specific, may provide the best arbiter of voice and speech performance and, possibly, rehabilitation success. That is, comprehensive perceptual judgments of alaryngeal speech may provide the ultimate index of postlaryngectomy communication effectiveness or success as perceived by the listener in a communication context (Doyle & Eadie, in press).

Given that alaryngeal voice is characterized by a substantial number of parameters, the interaction among these variables must be delineated so that a comprehensive approach to rehabilitation is undertaken. Therefore, auditory-perceptual judgments of "naturalness" and "severity" may provide two meaningful indicators of rehabilitation success because they represent a multidimensional, overall index of the voice/speech signal. If such auditory-perceptual features hold promise as dimensions that can be assessed in order to better define a given alaryngeal speaker's voice/speech character, we are then required to ascertain what method of auditory-perceptual evaluation is most appropriate. The essential question in this context centers specifically on whether the auditory-perceptual feature under evaluation is prothetic or metathetic. This obliges those interested in alaryngeal voice and speech to first address issues of what type of scaling procedure is most appropriate for each perceptual dimension and then determine the reliability of such measures as a clinical tool. As such, it is important that the validity of scaling multidimensional perceptual phenomena such as naturalness and overall severity be investigated in TE speakers. Consequently, the purpose of the present study was to (a) determine the psychophysical nature of speech naturalness and severity in TE speakers using Stevens' (1975) methods and (b) determine the construct validity of rating scales used for perceptual dimensions in this population of speakers.

Method

Participants

Speakers

The speakers who provided speech samples included 20 male adults who had undergone total laryngectomy and a TE puncture as a secondary procedure. Speakers ranged in age from 42 to 69 years old (M = 60.7 years) and were self-reported to be in good general health at the time of their participation. All speakers were at least 6 months post-TE puncture voice restoration, but none had used this method of alaryngeal speech for more than 5 years. All reported TE speech to be their primary method of verbal communication. All speakers used nonindwelling, low-profile Blom-Singer TE puncture voice prostheses (InHealth Technologies, Carpeneria, CA) at the time voice recordings were obtained.

The selection of voice samples was random but was made from a preselected set of samples rated as "better than average" based on perceptual judgments of experienced clinicians. All TE speaker samples were recorded using a unidirectional microphone (AKG-451-E) and a digital audiotape (DAT) research-quality player/recorder (Sony DTC-57ES) routed to an audio mixer (Teac 2-A). All TE speakers had bilateral pure-tone hearing within normal limits (< 30 dB) for the octave frequencies 250 to 2000 Hz and were native English speakers.

Listeners

Twenty graduate students in speech-language pathology served as listeners (age range, 22–34 years). All listeners were considered to be naive to voice pathology issues. Listeners reported no history of any hearing, speech, voice, or language difficulties.

Stimulus Tapes

Listening tapes were created by extracting the second sentence of Fairbanks' (1960) Rainbow Passage ("The rainbow is a division of white light into many beautiful colors") from samples produced by each of the speakers. The modulus was chosen from the 20 TE speaker samples by consensus judgment between a highly experienced clinician and a clinician with less than 5 years experience. The modulus voice sample was selected to represent the approximate midpoint for the features of naturalness and overall severity among the 20 speakers.

Four master listening DAT tapes (Sony DTC –57ES) were constructed: naturalness scaled with EAI (n-EAI), naturalness scaled with DME (n-DME), overall voice severity scaled with EAI (s-EAI), and overall voice severity scaled with DME (s-DME). Speaker order was randomized and there was a 5-second interstimulus interval between each pair of samples. Five stimulus samples were presented twice in each condition (n-EAI, n-DME, s-EAI, s-DME) to allow for assessment of intrarater reliability.

Listening Task

Listeners were required to evaluate each of the TE speakers' voice samples for naturalness and overall severity using both EAI and DME scaling procedures. Naturalness was defined as "a perceptually derived, overall description of prosodic adequacy." That is, speech was defined as "natural if it conforms to the listener's standards of rate, rhythm, intonation, and stress pattern, and if it conforms to the syntactic structure of the utterance being produced" (adapted from Yorkston, Beukelman, & Bell, 1988, p. 356). Overall severity of the voice was defined as "a comprehensive measure of how 'good' or 'poor' the voice sample is judged to be by the listener." This judgment was based on "multiple factors which ultimately range from normal to profoundly impaired on a continuum (e.g., normal, nearly normal, mild, mild-to-moderate, moderate, moderate-to-severe, severe, severe-to-profound, and profound)." The definitional markers represented each point on the 9-point EAI scale used for each of the naturalness and overall severity ratings. Listeners were divided into two groups of 10. Order of task presentation was randomized across listeners, and rating sessions were separated by 48 hours to control for learning effects.

Auditory-Perceptual Procedure

DME Scaling

To obtain auditory-perceptual judgments for the DME procedure, all listeners were first familiarized with the modulus sample and informed that it represented an arbitrary value of 100 on the DME scale. All judgments of experimental TE stimuli were made relative to the modulus. Listeners were instructed to scale speech samples that were twice as natural or severe at a value of "200"; TE speaker samples judged to be half as natural or severe were to be given a value of "50." The modulus was repeated after every 5 stimuli to prevent difficulty in recalling the modulus and causing a shift in the listeners' internal standard for the judgment in question (Kreiman et al., 1993). Listeners followed this procedure for both the naturalness and severity ratings.

EAI Scaling

Severity and naturalness ratings also were obtained using a 9-point EAI scale (Metz et al., 1990). For overall severity, listeners were asked to scale each sample according to the given definition, with a rating of "1" representing normal speech and "9" indicating the most severe. For naturalness ratings, "1" represented "very unnatural" speech and "9" indicated "very natural" speech.

Data Analysis

Stevens' (1975) psychophysical method of comparing DME and EAI ratings was used. Arithmetic means of the EAI scale values were plotted as a function of the geometric means of the DME values for all TE speech samples. This procedure was duplicated for both the ratings of severity and of naturalness. A linear relationship between the two sets of scaled judgments would indicate a metathetic continuum, whereas a downward-bowed, negatively accelerating curvilinear function would indicate a prothetic continuum.

Reliability

As noted, 5 (25%) of 20 speech samples were rated a second time for reliability purposes. Intrarater reliability was calculated using Pearson's correlation coefficients for DME conditions. Intrarater reliability values ranged from r = .619 to .979 for overall severity and from r = .499 to .992 for naturalness. For EAI ratings, values assigned by each rater were compared to values assigned to the repeated TE speaker samples. The percent agreement was calculated by counting the number of comparisons in which the assigned scale value did not differ by more than ±1 scale value, dividing this number by 5 for each speaker, and multiplying by 100. The same procedure was followed within ±2 scale values. The mean intrarater reliability was calculated for each condition by dividing the number of matches

by 100 (5 repeated samples × 20 raters) and multiplying by 100. The intrarater agreement ± 1 scale value for the n-EAI condition was 84%. The mean agreement ± 2 scale values was 96% for the n-EAI repeated judgments. The mean intrarater agreement ± 1 scale value for the s-EAI condition was 86%; within ±2 scale values, the mean agreement was 96%. Thus, intrarater reliability values were high, suggesting that each rater could maintain internal consistency in making naturalness and severity judgments.

The interrater reliability of the ratings was analyzed using Cronbach's alpha, which is equivalent to the intraclass correlation type (3, k) (Ebel, 1951; Shrout & Fleiss, 1979). For the DME measurements, the reliability of n-DME was .95 and the reliability of s-DME was .87. For interval ratings, the reliability of n-EAI was .96 and the reliability of s-EAI was .97. These results are consistent with previously reported ratings for other variables such as stuttering severity and speech naturalness (Martin, Haroldson, & Triden, 1984; Metz et al., 1990; Schiavetti et al., 1983). Because scaling data such as naturalness and severity are usually employed as group mean ratings for research purposes, the group reliability coefficients are more descriptive than the reliability of individual raters. Thus, care must be employed when these results are generalized to individual clinical judgments (Metz et al., 1990).

Results

Naturalness

EAI means were plotted as a function of the DME geometric means for the naturalness ratings of TE speakers (see Figure 1). Results indicated that the relationship was significantly predicted with the best-fit linear equation ($y = 0.0319x + 1.269$) for EAI naturalness values from DME estimates [$r^2 = .853$, $F(1, 18) = 104.717$, $p < .01$]. A test for curvilinearity revealed no significant improvement in the variance accounted for by any of the curvilinear models over the linear model. Visual inspection of the figure also revealed a good approximation to the raw data by the linear regression, with no apparent downward bowing.

Severity

For severity ratings, EAI means plotted as a function of the DME geometric means revealed a statistically significant result [$r^2 = .715$, $F(1, 18) = 45.173$, $p < .01$] for the curvilinear model as shown in Figure 2 (line of best fit: $\hat{y} = 4.752\text{Ln}(x) - 15.478$). This indicates that the curvilinear model accounted for a statistically significant amount of the variance observed, above and beyond that accounted for by a simple linear model.

Figure 1. Mean interval scale values of speech naturalness plotted as a function of geometric mean DME estimates for TE speakers.

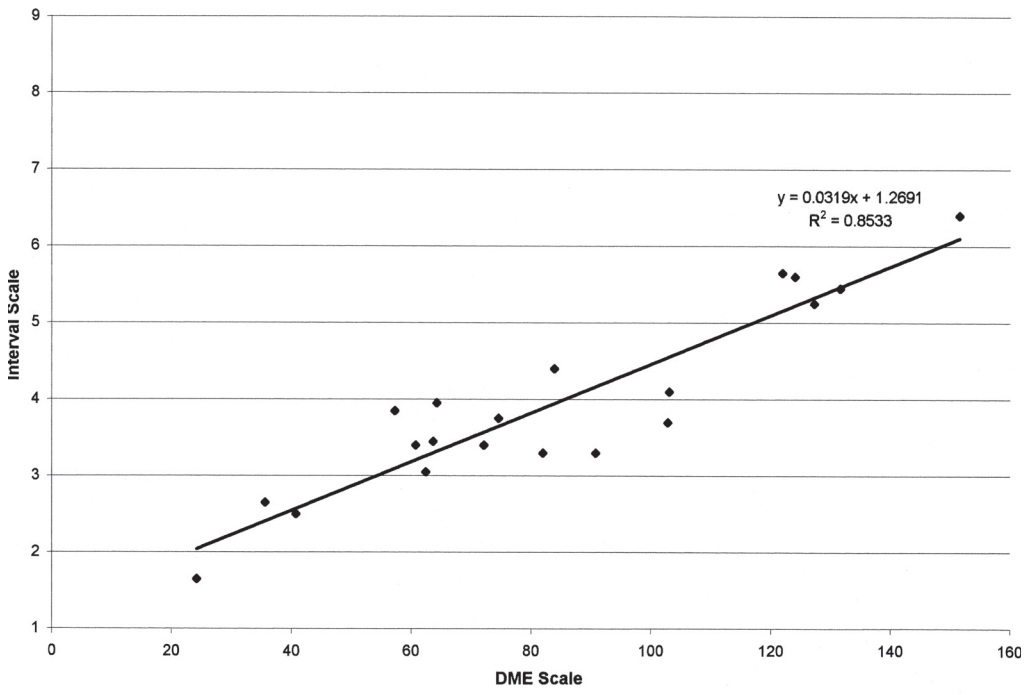
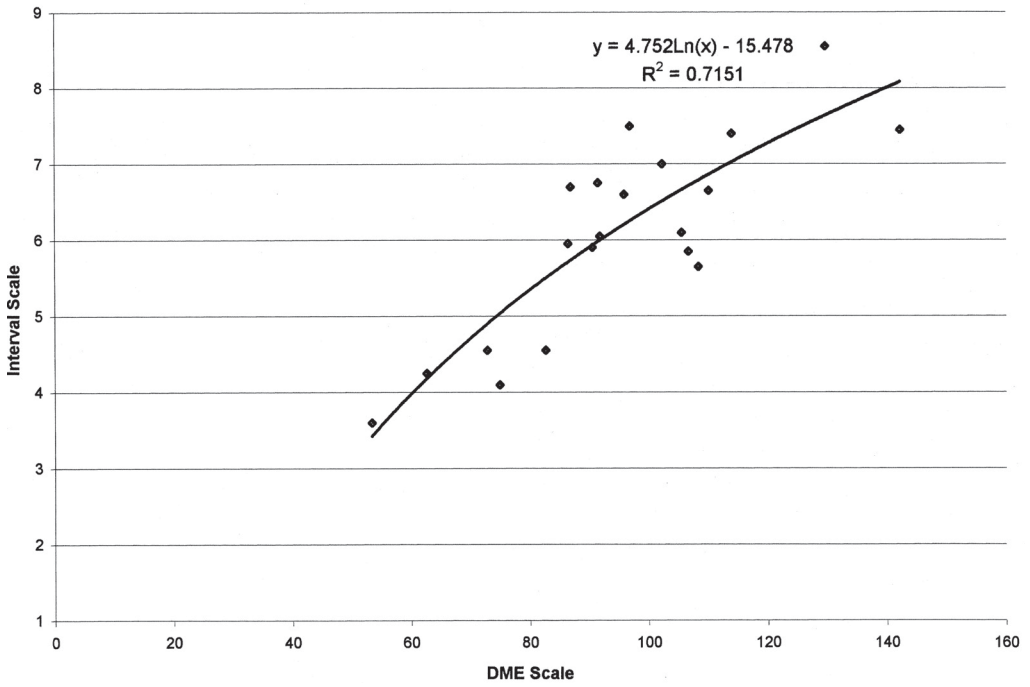


Figure 2. Mean interval scale values of voice/speech severity plotted as a function of the geometric mean DME estimations for TE speakers.



Discussion

The significant linear relationship between the EAI and DME naturalness ratings indicates that a metathetic continuum best represents the perceived naturalness of TE speakers' samples (variance for 85.3% of the metathetic continuum is accounted for). These results agree with findings for laryngeal speakers with and without stuttering impairments (Metz et al., 1990). Our results indicate that either EAI or DME scales can be used to validly measure speech naturalness in TE speakers. These results also suggest that the nature (i.e., prothetic vs. metathetic) of the dimension "naturalness" applies to both laryngeal speakers (e.g., Metz et al.) and the current alaryngeal speakers who use TE speech/voice. Given the strong relationship of naturalness to speech rate in prior speech pathology literature, it seems that the TE speakers' access to pulmonary air serves to retain the psychophysical nature of this feature for listener judgments despite clear differences in the voice "quality" of the TE signal. As such, we believe it confirms the ability of listeners to make independent assessments of naturalness in the presence of competing, and in the present case, clearly abnormal characteristics of the voice signal under evaluation. Based on this finding, the auditory-perceptual feature of "naturalness" appears to be a very robust dimension that can be applied consistently to a variety of voice and speech disorders (e.g., dysphonia, dysarthria, and TE speech). However, although the present data reveal consistency with previous studies of other disordered populations, further investigation is required to determine whether these results can be generalized to other populations of alaryngeal speakers (e.g., esophageal and artificial laryngeal speech).

In contrast to our findings for naturalness, a significant curvilinear relationship was found between the EAI and DME ratings of severity, indicating that a prothetic continuum best represents this dimension for the present TE speaker samples (variance for 71.5% of the prothetic continuum is accounted for). These results are the first to be reported in this area and suggest that the "severity" of TE speech, like stuttering severity, must be measured with DME scales to avoid the linear partitioning assumed by EAI scales. That is, using EAI scales to rate the severity of TE speakers' voice/speech introduces listener bias as they attempt to partition the perception of severity into equal intervals. Stevens (1974) indicated that with prothetic dimensions, listeners do not perceive intervals as equal at different locations along the scale. For example, the difference between the overall severity rating of "1" and "2" may not represent the same magnitude as the difference between "4" and "5" or between "8" and "9." Hence, from a clinical perspective, using EAI scales to measure intervention outcomes in relation to

"severity" is inappropriate, as it is difficult to interpret relative differences between scale values (i.e., a change from 1 to 2 may mean something different than a change from 5 to 6).

Wuyts, De Bodt, Molenberghs, Remacle, Heylen, Millet, et al. (2000) suggested that a global, comprehensive measure such as "severity" may correlate most highly with both objective measures and functional outcome/quality of life measures for a dysphonic speaker. Regardless of the disorder—laryngeal-based or alaryngeal—changes in "severity," a multidimensional perceptual dimension, would seem to be the most meaningful to the client, as this measure may correspond directly to the impact of the disorder on daily activities and one's participation in society. Because of this, overall severity may have clinical utility as a measure of TE speech proficiency provided it is scaled using the appropriate scaling method in a consistent, structured manner.

Over the past decade, perceptual evaluation of voice and voice disorders has gained considerable attention (cf. Kent, 1996; Kreiman et al., 1993; etc.). However, although the focus of research has been directed toward disordered "laryngeal" voice, less attention has been directed at the individual who uses an alaryngeal communication method (Doyle & Eadie, in press). Much of the recent work addressing voice perception has indicated that numerous factors must be considered in relation to interpreting the data obtained. Factors such as listener experience, definitions of the feature under assessment, scale resolution, interactions between the task and the listener, and so on, are elements to consider when perceptual questions are critically addressed (Kreiman et al., 1993). Additionally, before the current investigation, no information exists on the potential influence that an abnormal and "non-normal" voice source has on a listener's judgment of auditory-perceptual features. This factor would seem to have critical importance on how we conduct and interpret studies of alaryngeal voice and speech. Of course, if the validity of the scale is under question, then these factors are moot considerations (cf. Gescheider, 1988).

Multidimensional factors inherent in the alaryngeal speech signal must be considered of primary importance from the standpoint of rehabilitation outcomes. For example, the ultimate arbiter of alaryngeal speech "success" likely finds some component in the realm of the degree to which the individual speaker is recognized as being "abnormal" or different from normal expectation (Van Riper, 1978). Loss of the larynx and use of a non-normal voicing source certainly increase the likelihood that the speech signal will call attention to itself. These considerations extend beyond issues of postlaryngectomy voice/speech acquisition (regardless of alaryngeal mode), speech intelligibility, and the considerable attention given to unidimensional measures of the voice signal (e.g., pitch, loudness, and rate) as a sole measure of alaryngeal speech success and/or proficiency (Doyle & Eadie, in press). We believe that efforts directed toward more comprehensive methods of describing the composite nature of alaryngeal speech have clinical value. In this regard, more global perceptual features may best define the TE speaker's ability in comparison to other TE speakers. Multidimensional indicators of success may be found in both severity and naturalness judgments. Thus, it is vital that perceptual measures of these important dimensions of TE speech are validly scaled, because they could potentially serve as meaningful indicators of rehabilitation success.

The present data suggest that naturalness and overall severity, if assessed appropriately, may provide reliable and valid clinical measures of TE speakers. However, the clinical literature on alaryngeal speech is replete with auditory-perceptual dimensions, many of which are poorly defined. This raises new questions about the need to more specifically define and evaluate specific attributes that may define the perceptual character of TE, as well as other methods of alaryngeal speech. Further, as exploration of such attributes occurs, researchers must carefully assess the nature of any given attribute (i.e., is the feature under study prothetic or metathetic?) in order for it to be assessed in a valid manner. Findings from the present study also suggest that further empirical evaluation of the auditory-perceptual dimension of alaryngeal voice and speech dimensions using Stevens' (1975) procedures may ultimately lead to better understanding of the perceptual correlates of postlaryngectomy voice and speech. Other perceptual dimensions that have been reported in the literature, such as acceptability and pleasantness, might also be assessed to determine whether they represent prothetic or metathetic continua. Consequently, results from these evaluations may lead to development of potential outcome measures following postlaryngectomy voice restoration. Ultimately, implications for future research suggest the development of a clinical scale that uses DME measurement in order to avoid applying the incorrect type of scale for the attributes measured (Schiavetti, 1984). We are currently exploring these questions and others in order to provide an improved understanding of the auditory-perceptual nature of postlaryngectomy communication.

Acknowledgments

This paper was written in partial fulfillment of the thesis requirements for obtaining a PhD in Rehabilitation Sciences at the University of Western Ontario. The authors wish to acknowledge the continued support of the Harmonize for Speech Fund, the Ontario Barbershop Association, and the Canadian Institutes of Health Research (CIHR).

References

Barry, S. J., & Kidd, G. (1981). Psychophysical scaling of distorted speech. *Journal of Speech and Hearing Research, 24*, 44–47.

Blom, E., Singer, M., & Hamaker, R. (1986). A prospective study of tracheoesophageal speech. *Archives of Otolaryngology–Head and Neck Surgery, 112*, 440–447.

Blom, E. D., Singer, M. I., & Hamaker, R. C. (1998). *Tracheoesophageal voice restoration following total laryngectomy*. San Diego, CA: Singular.

Doyle, P. C. (1994). *Foundations of voice and speech rehabilitation following laryngeal cancer*. San Diego, CA: Singular.

Doyle, P. C., & Eadie, T. L. (in press). The perceptual nature of alaryngeal voice and speech. In P. C. Doyle & R. L. Keith (Eds.), *Contemporary considerations in the treatment and rehabilitation of head and neck cancer*. Austin, TX: Pro-Ed.

Ebel, R. L. (1951). Estimation of the reliability of ratings. *Psychometrika, 16*, 407–424.

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

Finizia, C., Dotevall, H., Lundström, E., & Lindström, J. (1999). Acoustic and perceptual evaluation of voice and speech quality. *Archives of Otolaryngology–Head and Neck Surgery, 125*, 157–163.

Gerratt, B. R., Till, J., Rosenbek, J. C., Wertz, R. T., & Boysen, A. E. (1991). Use and perceived value of perceptual and instrumental measures in dysarthria management. In C. A. Moore, K. M. Yorkston, & D. R. Beukelman (Eds.), *Dysarthria and apraxia of speech* (pp. 77–93). Baltimore: Paul H. Brookes.

Gescheider, G. A. (1988). Psychophysical scaling. *Annual Review of Psychology, 39*, 169–200.

Hillman, R. E., Walsh, M. J., Wolf, G. T., Fisher, S. G., & Hong, W. K. (1998). Functional outcomes following treatment for advanced laryngeal cancer. *Annals of Otology, Rhinology and Laryngology, 107*, 2–27.

Keith, R. L., & Darley, F. L. (1986). *Laryngectomee rehabilitation* (2rd ed.). San Diego, CA: College-Hill Press.

Keith, R. L., & Darley, F. L. (1994). *Laryngectomee rehabilitation* (3rd ed.). Austin, TX: Pro-Ed.

Kent, R. (1996). Hearing and believing: Some limits to the auditory perceptual assessment of speech and voice disorders. *American Journal of Speech-Language Pathology, 5*(3), 7–23.

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

Martin, R. R., Haroldson, S. K., & Triden, K. A. (1984). Stuttering and speech naturalness. *Journal of Speech and Hearing Disorders, 49*, 53–58.

Metz, D. E., Schiavetti, N., & Sacco, P. R. (1990). Acoustic and psychosocial dimensions of the perceived speech naturalness of nonstutterers and posttreatment stutterers. *Journal of Speech and Hearing Disorders, 55*, 516–525.

Moon, J. B., & Weinberg, B. (1987). Aerodynamic and myoelastic contributions to tracheoesophageal voice production. *Journal of Speech and Hearing Research, 30*, 387–395.

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

Pindzola, R. H., & Cain, B. H. (1989). Duration and frequency characteristics of tracheoesophageal speech. *Annals of Otology, Rhinology and Laryngology, 98*, 960–964.

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

Schiavetti, N. (1984). Scaling procedures for quantification of speech, language, and hearing variables. In R. G. Daniloff (Ed.), *Articulation assessment and treatment issues* (pp. 237–253). San Diego, CA: College-Hill.

Schiavetti, N., Metz, D. E., & Sitler, R. W. (1981). Construct validity of 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., & Sitler, R. W. (1983). Direct magnitude estimation and interval scaling of stuttering severity. *Journal of Speech and Hearing Research, 26*, 568–573.

Sewall, A., Weglarski, A., Metz, D. E., Schiavetti, N., & Whitehead, R. L. (1999). A methodological control study of scaled vocal breathiness measurements. *Contemporary Issues in Communication Sciences and Disorders, 26*, 168–172.

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

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

Snidecor, J. C. (1978). *Speech rehabilitation of the laryngectomized* (2nd ed.). Springfield, IL: Charles C. Thomas.

Stevens, S. S. (1974). Perceptual magnitude and its measurement. In E. C. Caterette & M. P. Friedman (Eds.), *Handbook of perception* (Vol. 2, pp. 22–40). New York: Academic Press.

Stevens, S. S. (1975). *Psychophysics: Introduction to its perceptual, neural, and social prospects*. New York: Wiley.

Tardy-Mitzell, S., Andrews, M., & Bowman, S. A. (1985). Acceptability and intelligibility of tracheoesophageal speech. *Archives of Otolaryngology, 111*, 212–215.

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

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

Van As, C. J., Hilgers, F. J. M., Verdonck-de Leeuw, I. M., & Koopmans-van Beinum, F. J. (1998). Acoustical analysis and perceptual evaluation of tracheoesophageal prosthetic voice. *Journal of Voice, 12*, 239–248.

Van Riper, C. (1978). *Speech correction: Principles and methods.* Englewood Cliffs, NJ: Prentice Hall.

Whitehill, T. L., Lee, A. S. Y., & Chun, J. C. (2002). Direct magnitude estimation and interval scaling of hypernasality. *Journal of Speech, Language, and Hearing Research, 45*, 80–88.

Williams, S., & Watson, J. B. (1985). Differences in speaking proficiencies in three laryngectomy groups. *Archives of Otolaryngology, 111*, 216–219.

Wuyts, F. L., De Bodt, M. S., Molenberghs, G., Remacle, M., Heylen, L., Millet, B., et al. (2000). The Dysphonia Severity Index: An objective measure of vocal quality based on a multiparameter approach. *Journal of Speech, Language, and Hearing Research, 43*, 796–809.

Yorkston, K. M., Beukelman, D. R., & Bell, K. R. (1988). *Clinical management of dysarthric speakers*. Boston: College-Hill Press.

Zraick, R. I., & 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.

Received March 18, 2002

Accepted June 17, 2002

DOI: 10.1044/1092-4388(2002/087)

Contact author: Tanya L. Eadie, MSc, Voice Production Laboratory, School of Communication Sciences and Disorders, University of Western Ontario, London, Ontario N6G 1H1, Canada. E-mail: teadie@uwo.ca