Influence of Speaker Gender on Listener Judgments of Tracheoesophageal Speech

*Tanya L. Eadie, †Philip C. Doyle, Kerry Hansen, and †Paul G. Beaudin

*Seattle, Washington and †London, Ontario, Canada

Summary: The objectives of this prospective and exploratory study are to determine: (1) naïve listener preference for gender in tracheoesophageal (TE) speech when speech severity is controlled; (2) the accuracy of identifying TE speaker gender; (3) the effects of gender identification on judgments of speech acceptability (ACC) and naturalness (NAT); and (4) the acoustic basis of ACC and NAT judgments. Six male and six female adult TE speakers were matched for speech severity. Twenty naïve listeners made auditory-perceptual judgments of speech samples in three listening sessions. First, listeners performed preference judgments using a paired comparison paradigm. Second, listeners made judgments of speaker gender, speech ACC, and NAT using rating scales. Last, listeners made ACC and NAT judgments when speaker gender was provided coincidentally. Duration, frequency, and spectral measures were performed. No significant differences were found for preference of male or female speakers. All male speakers were accurately identified, but only two of six female speakers were accurately identified. Significant interactions were found between gender and listening condition (gender known) for NAT and ACC judgments. Males were judged more natural when gender was known; female speakers were judged less natural and less acceptable when gender was known. Regression analyses revealed that judgments of female speakers were best predicted with duration measures when gender was unknown, but with spectral measures when gender was known; judgments of males were best predicted with spectral measures. Naïve listeners have difficulty identifying the gender of female TE speakers. Listeners show no preference for speaker gender, but when gender is known, female speakers are least acceptable and natural. The nature of the perceptual task may affect the acoustic basis of listener judgments.

Key Words: Tracheoesophageal speech—Listener judgments—Acoustic measures.

Accepted for publication August 14, 2006.

Portions of this paper were presented at the Voice Foundation's 35th Annual Symposium: Care of the Professional Voice, Philadelphia, PA, June 2006.

From the *Department of Speech and Hearing Sciences, University of Washington, Seattle, Washington; and the †Doctoral Program in Rehabilitation Sciences, University of Western Ontario, London, Ontario, Canada.

Address correspondence and reprint requests to Tanya L. Eadie, Department of Speech and Hearing Sciences, University of Washington, 1417 NE 42nd Street, Seattle, WA. E-mail: teadie@u.washington.edu

Journal of Voice, Vol. 22, No. 1, pp. 43–57
0892-1997/$34.00

doi:10.1016/j.jvoice.2006.08.008

INTRODUCTION

Tracheoesophageal (TE) puncture voice restoration has become an increasingly common method of postlaryngectomy voice and speech rehabilitation over the past two decades.[1,2] In the TE puncture procedure, a fistula is surgically created between the trachea and the esophagus. When air is exhaled and the tracheostoma is occluded, pulmonary air is shunted through a one-way TE puncture voice prosthesis into the esophageal reservoir; this air can then be used to create oscillation of the pharyngoesophageal (PE) segment which serves as the alaryngeal source of postlaryngectomy voice.[1,3] The TE voicing source is characterized by considerable deviance when compared to the normal voice and marked by substantial variability between speakers.

The variability in TE voice and speech is observed for unidimensional physical measures of frequency, pause time, the magnitude of aperiodic noise, etc.[3] Although acoustic data indicate that TE speech closely approximates normal laryngeal voice with regard to frequency,[4–8] intensity,[6,7,9] and duration,[6,7,9–11] the TE signal is clearly perceived as abnormal by the listener.[12] To date, however, concerns related to distinctions between male and female alaryngeal voices have been limited. Given perceptual expectations of the voice signal specific to gender, the impact of TE voice and speech on the acoustic and, therefore, auditory-perceptual end-product remain of considerable interest.

In one of the few studies of female TE speakers, Trudeau and Qi[8] found that the acoustics of female TE speech closely paralleled the characteristics of male TE speech for reading rate, number of pauses, fundamental frequency (F_0) in reading, and for directional perturbation (jitter and shimmer). One expected difference between male and female speakers was anticipated for F_0; however, Trudeau and Qi[8] found that the F_0 of female TE speakers in their study corresponded to a range at or below that of normal laryngeal male adults. Instead, they found that the greatest difference between female TE voices and male TE voices from past studies was related to greater perturbation for females. More recently, Bellandese et al[4] found that ‘‘excellent’’ female TE speakers differed from female laryngeal speakers for mean F_0 for vowels, signal-to-noise ratio (SNR), and duration for a reading passage, number of pauses, and syllables per minute. These results indicated that F_0 for ‘‘excellent’’ female TE speakers may be closer to that expected for normal females. These results also correspond with suggestions of the potential for a ‘‘feminine’’ esophageal voice that are related to increased pitch, faster rate, and better voice quality.[13,14] Thus, questions pertaining to how a listener perceives the alaryngeal voice in the context of one's gender may be raised with substantial rehabilitation implications.

Historically, the concept of voice and speech ‘‘effectiveness’’ has been used in conjunction with ‘‘speech acceptability’’ as a means of alaryngeal speaker categorization.[15–17] However, the success of postlaryngectomy speech rehabilitation is more likely determined by how much the new voice is judged by the listener to deviate from normal expectation, at least in relation to dimensions such as overall severity, acceptability (ACC), and naturalness (NAT).[18,19] Dimensions such as these have been used in past studies to evaluate the success of medical treatments and surgical procedures associated with laryngeal cancer, to compare across groups of alaryngeal speakers, and to evaluate speech success within a group.[5,16,20–22] In general, results have shown that regardless of gender or perceived communicative ‘‘excellence,’’ TE speakers are judged as having voices that are less acceptable and poorer in voice quality than normal speakers.[12,20,21]

Recently, Eadie and Doyle[22] investigated auditory-perceptual dimensions in TE speakers. Using psychophysical scaling methods, 15 naïve listeners evaluated 28 TE speech samples (22 males and 6 females) for overall speech severity, NAT, ACC, and pleasantness. In this study, listeners were provided information on each speaker's gender. Results showed that listeners discriminated among TE speech samples relative to these auditory-perceptual dimensions and that male speakers were judged as having significantly better, more acceptable, and more pleasant voices than women. Male and female speakers, however, were not different relative to speech NAT, a feature that primarily focuses the rate of one's speech. These data suggested that women who use TE speech may be differentially penalized for particular perceptual dimensions.

The findings reported by Eadie and Doyle[22] for female TE speakers could result from a number of factors. First, because speaker gender was provided to listeners, it is unclear whether there was something about the composite ''quality'' of either the male or female TE speaker signals that inherently resulted in listeners preferring males over females. In this regard, it is unknown how accurate listeners would have been in judging speaker gender had they not been provided with this information *a priori*.

Weinberg and Bennett[17,23] investigated listeners' abilities to identify gender in a group of esophageal speakers and found consistent differences in how they distinguished males from females; this included identification of F_0 levels that clearly resulted in masculinization of the female esophageal voice. However, despite the noted acoustic changes, listeners still identified speaker gender with a high degree of accuracy, although correct identification of male speakers was 20% higher than for females.[23] In a more recent study, Searl and Small[24] examined naïve listeners' abilities to identify gender from six female and six male TE speech samples. Results indicated that listeners accurately identified female speakers 89.5% of the time, and male speakers were identified with 88.6% accuracy. In addition, listeners were asked to judge the masculinity and femininity of the speech samples and found a moderate relationship for male speakers, but a weaker relationship for females. That is, although listeners were able to accurately identify female TE speakers, listeners' perceptions of femininity were not strongly related to these judgments. Although these results suggest that TE speaker gender is equally and accurately identified by naïve listeners, one must consider the speech samples used in the study. Speech samples were those judged as average to excellent by two experienced speech-language pathologists. As a result, these data might not generalize to TE speakers who are less proficient and more representative of ''typical'' TE speakers.

A second reason that naïve listeners might have judged male TE speakers more favorably than females in the study by Eadie and Doyle[22] may relate to how listeners performed these judgments. Typically, listeners make auditory-perceptual judgments by comparing the presented speech signal with an internal referent, or a prototype of what is expected for the dimension being evaluated.[25] These internal referents are derived from a listener's experience. Because a listener's internal perceptual referent is almost certainly formed relative to laryngeal speakers, the unusual nature of TE speech and information related to speaker gender may have led listeners to use a ''normal'' comparative template. Thus, listeners might have penalized female TE speakers more than males because female voice signals may have differed more from the ''typical'' female laryngeal signal than male TE speakers differed from their laryngeal counterparts.[25] The differential penalty identified by Eadie and Doyle[22] for female TE speakers could also relate to the underlying acoustic variables used by listeners to make judgments of ACC, severity, and pleasantness. Because there were strong relationships among these perceptual variables, listeners could have used similar cues to make their judgments.

Past explorations of auditory-perceptual correlates underlying assessments of postlaryngectomy speech have been related to acoustic variables and levels of source aperiodicity.[3] Acoustic variables that have been associated with better overall ACC ratings for male esophageal speakers include greater intensity, higher F_0, faster rate, and improved fluency[16,26,27] with similar variables identified for the male TE speech signal.[28] Obviously, reductions in F_0 have considerable potential to be of greater limitation to females than males when one considers that the typical (tracheo)esophageal voicing source has frequently been reported to be well below 100 Hz.[7,15,29] In fact, it has commonly been acknowledged that females who use an ''esophageal'' voicing source (whether it be traditional esophageal or TE) will find that level to be approximately one octave lower than normal expectations.[3] For example, in the study by Trudeau and Qi,[8] the average F_0 level was found to be 108.6 Hz. While the influence of F_0 is primary as a perceptual index of speaker gender, multidimensional aspects of the speech signal cannot be discounted in relation to the perceptual relevance on gender identification in those who use an alaryngeal mode of communication. Although these relationships have been investigated in male alaryngeal signals, such relationships have not been investigated in female alaryngeal speakers. However, investigating the perceptual and acoustic basis underlying judgments of women in this context, and for TE speech in particular, has become increasingly important due to the proportion of women diagnosed with laryngeal cancer[30] and those who will undergo TE puncture voice restoration. Results have implications for both assessment and treatment of individuals who use this mode of alaryngeal speech, and ultimately, the long-term outcomes of rehabilitation success. Therefore, the purpose of the current study sought to determine the relative contributions of a number of factors that could potentially lead to differential penalty of women's TE voices in perceptual judgments. In doing so, it was necessary to control variables such as overall speech severity and *a priori* listener knowledge of speaker gender. Consequently, the present series of studies were designed to answer the following experimental questions:

(1) Do differences exist in naïve listeners' preference of male and female TE speakers when gender is unknown to listeners and when TE speech severity is controlled?
(2) Does the accuracy of gender identification of TE speakers by naïve listeners differ when speech severity is controlled?
(3) What is the relative contribution of known gender to listeners' evaluations of NAT and ACC of TE speech?
(4) What acoustic parameters underlie naïve listeners' judgments of NAT and ACC of male and female TE speech when gender is unknown and known?

METHOD

Participants

Speakers

Audio recordings from 12 TE speakers (6 males and 6 females) were selected from speaker samples used in the study by Eadie and Doyle.[22] In that study, 15 naïve listeners judged the overall severity for each speaker sample using direct magnitude estimation (DME) scaling procedures. Listeners were provided information related to the gender of each speaker, and they were asked to judge each speaker sample relative to modulus samples that were arbitrarily valued at 100. The modulus samples were selected to represent those samples in the midpoint of the severity scale.[22] To select samples in the present study, the geometric mean for each speaker was examined from the prior study. Each female speaker was subsequently matched with a male speaker who had an approximately equal speech severity score and a similar age. Severity ratings and demographic information are provided in Table 1. The mean age for male speakers was 69 years (age range, 59–81 years), and 65 years (age range, 57–72 years) for female speakers. The mean speech severity ratings for males and females were 79 and 83, respectively. Mean time postlaryngectomy for males was 126 months and mean time using TE speech as the primary mode of communication was 96 months; females were at an average of 95 months postlaryngectomy and had used TE speech for 92 months. All participants reported using TE speech as their primary mode of communication, all had received radiation treatment, and all were English speakers. Speakers were excluded if they

TABLE 1. *Demographic Characteristics of the Speakers, Including Age in Years, Gender, Time Postlaryngectomy in Months, Time Using TE Speech as Primary Mode of Communication in Months, and DME Speech Severity Ratings*

Speakers	Age (y)	Gender	TPL	Time TE	Speech Severity
1	63	F	234	234	16.1
2	69	M	252	144	22.3
4	71	M	31	24	28.9
6	72	F	93	93	36.9
13	63	F	108	102	66.9
14	59	M	145	144	72.1
16	64	F	24	12	93.4
18	81	M	180	120	110.4
19	76	M	90	90	115.7
21	62	M	60	54	124.9
23	71	F	60	60	136.4
24	57	F	48	48	146.6

Note: a smaller score indicates a more severe speech score.
Abbreviations: M, male; F, female; TPL, time postlaryngectomy; time TE, time using TE speech as primary mode of communication.

reported any history of other medical conditions that would affect speech, language, or hearing.

Listeners

Twenty (17 females and 3 males) undergraduate and graduate students (mean age = 22 years, range, 19–27 years) served as listeners in this study. These listeners were considered naïve to voice pathology issues and had no prior experience with or exposure to alaryngeal speech. All were screened for bilateral pure-tone hearing at 25 dB for the octave frequencies 250 to 4000 Hz, and were native English speakers. Permission to conduct this study was formally granted by the Human Subjects Committee at the University of Washington.

Speech stimuli

All speakers provided recordings of a standard reading passage used in voice analysis, Fairbanks' Rainbow Passage (Fairbanks, 1960).[31] Speech samples were recorded using a headset microphone (Shure SM10a; Shure Incorporated, Niles, IL) and a digital minidisk (MD) research quality recorder (Sony MZ-R55; Sony Corp., New York, NY) or a digital audiotape portable recorder (Sony PCM-M1) in a quiet background with minimal background noise. Both sets of recordings were digital originals and were recorded at a sampling rate of 48 kHz.

The digital recordings were transferred to a personal computer and saved as WAV files using acoustic software (*Sony Soundforge 7.0*). Each sample was edited to include the second sentence of the Rainbow Passage, ''The rainbow is a division of white light into many beautiful colors.'' All samples were then entered into a perceptual software program (*Win/Ecos*)[32] that randomizes speaker presentation, and permits listener responses using a prescribed scale.

Listening procedures

Listening session 1: paired comparison

The first listening session used the method of paired comparison and an accompanying 60-mm visual analog scale (VAS). Each TE speaker sample was paired with every other speaker sample in both A-B and B-A conditions (n = 132 speaker pairs). Twenty-eight speaker pairs were repeated for intrarater reliability (N = 160 speaker pairs). Each listener was seated at a computer and was presented speaker samples over headphones. Before making judgments, listeners were familiarized with the scale and provided instructions. Listeners were asked to use a mouse-controlled VAS to rate which speaker of each pair they preferred, with one end of the scale labeled ''prefer speaker 1'' and the other end labeled ''prefer speaker 2.'' No definition of preference was provided for making judgments.

Listening session 2: gender identification, ACC, and NAT ratings

The second listening session was separated from the first session by 1 week. Before making any speaker judgments, listeners were first familiarized with the gender rating scale.[24] This was a 60-mm semantic differential VAS with end points labeled as ''definitely male'' (''0'' on the scale) and ''definitely female'' (''60'' on the scale). Compared to a binary choice (male/female), the 60-mm VAS allowed for a finer description of listeners' gender identification by embedding a confidence interval into the scale itself. For example, a rating between 0 and 15 mm suggests that listeners were quite clear the voice was ''male,'' between 16 and 30 mm is indicative of a less definitive or ''somewhat male'' rating, between 31 and 45 mm indicates a neutral or ''somewhat female'' rating, and between 46 and 60 mm indicates a definitive perception of ''female.''

After making judgments using the gender rating scale, listeners were asked to make judgments of speech ACC or NAT using a 60-mm VAS. The order of these judgments was counterbalanced and speaker order was randomized for each listener. In making judgments about speech ACC, 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?''[20] The end points of the scale were labeled as 0 = ''Very Unacceptable'' to 60 = ''Very Acceptable.'' Speech samples were defined as ''natural if [they] conform to the listener's standards of rate, rhythm, intonation, and stress pattern''.[33(p.356)] The end points of the scale were labeled as 0 = ''very unnatural'' to 60 = ''very natural.'' These definitions were identical to those used in the study by Eadie and Doyle.[22] Because the results differed between ACC and NAT in that prior study,[22] it was anticipated that listeners in the present study could differentiate between these dimensions. Eight speaker samples were randomly repeated for each of the gender ratings, ACC ratings, and NAT ratings ($n = 20$ speech samples per scale) to calculate intrarater reliability.

Listening session 3: ACC and NAT ratings with gender identified

Listening session 3 was separated from listening session 2 by 1 week to help control for potential learning effects. Listeners were again asked to make judgments of the speaker samples for ACC and NAT, using the same procedures used during session 2. Order of the rating tasks was counterbalanced across listeners. In this rating session, listeners were provided with the speaker gender (indicated by an ''M'' or ''F'' on a form which coincided with the speaker order on the computer). Listeners were then asked to make judgments relative to what would be expected for a speaker of the same gender for both ACC and NAT. Eight speech samples were repeated to calculate intrarater reliability ($N = 20$ for each rating dimension).

Acoustic analysis

The acoustic analyses were performed using *Computerized Speech Lab* (CSL) with a 4500 external module.[34] Speaker samples were downsampled to 10 kHz and low-pass filtered at 1000 Hz. Pitch-related measures were derived using the running speech analysis of CSL after visual inspection of the narrowband spectrogram for periodic voiced segments. Measures included mean F_0, SD F_0, min F_0, and max F_0.

Total duration of the sentence, the number of pauses, duration of pauses, and syllables per second were calculated. By placing time cursors at the beginning and end of each sentence sample, the duration was determined. Syllables per second were calculated by dividing the number of syllables in the sentence (20) by the duration of the sentence (in seconds). The number and duration of pauses was determined from the waveform tracings; a pause was considered to occur when the tracing returned to baseline for at least 200 milliseconds.[7]

Spectral measures were determined using the *interactive voice analysis system* (IVANS).[35] Five parameters were extracted from the continuous speech data for linear regression prediction of ACC and NAT ratings. These parameters included long-term average speech (LTAS) spectrum measures of: (1) overall spectral tilt (SPTILT) and (2) voiced frames of spectral tilt (TILT-V), and linear-prediction (LP) modeling measures of (3) LP-SNR, (4) pitch amplitude (PA), and (5) spectral flatness ratio (SFR). LP-based measures were selected because they do not require highly precise estimates of the F_0 contour, which is often difficult to extract from aperiodic signals (ie, alaryngeal voices). LTAS spectrum measures (eg, spectral tilt) were selected because they do not rely on the extraction of ''voiced'' speech as is necessary in LP modeling–based measures.[36]

Two LTAS measures were computed from the continuous speech data. For the first measure, the input speech was segmented into 30-millisecond frames. Each frame was then partitioned using a Hamming window and transformed into the spectral domain. Individual frame spectra were then averaged to obtain the LTAS, and the spectral tilt was calculated as the mean dB difference per octave. For the second LTAS measure, only the voiced frames were used in the computation.

The next three measures were based on linear prediction modeling. The first measure was LP-SNR.[36] The autocorrelation method was used to estimate the LP model coefficients. The inverse-filtered speech then had long-term correlations removed by a second LP model. The SNR estimate was calculated from the ratio of energies of the input speech signal and the residue of the second LP model. PA and SFR parameters were calculated for voiced frames using only the voice detection algorithm.[36]

Data analysis

Paired comparison

Data from each listener were converted into a mean ''preference'' score for each speaker collapsed across speaker order (11 speaker pairs $\times$ 20 listeners = 220 judgments per speaker). A smaller score on the VAS indicated that the listeners preferred that speaker; a bigger score on the VAS indicated that listeners preferred the matched pair (ie, the other speaker). Speakers also were ranked according to their mean preference scores; matched paired t tests were then performed between male and female speakers using these scores.

Reliability for paired comparison

Twenty-eight speaker pairs were repeated randomly throughout the paired comparison procedure for reliability purpose. The relationship between initial and repeated samples was calculated using Pearson correlation coefficients. Intrarater reliability ranged from $r = .47$ to $r = .94$ (mean $r = .71$). Interrater reliability was calculated by examining the relationships among listeners' ratings for the speaker pairs using intraclass correlation coefficients (ICCs) (group averages model). The ICC value was .946 for the paired comparison preference judgments.

Gender identification, NAT, and ACC ratings

To evaluate listener identifications of TE speaker gender, the number of correctly identified speakers (out of six) was calculated from listener responses on the gender rating scale. Ratings of 0 to 15 mm were considered "definitely male," 16 to 30 mm were considered "somewhat male", 31 to 45 mm were "somewhat female," and 46 to 60 mm were "definitely female." Therefore, accurate responses for males included scores of 0 to 30 mm; speakers were identified females if scores were from 31 to 60 mm.

The average of all listeners' judgments was determined for each speaker for speech ACC and NAT. These averages were calculated when gender was unknown/not provided (listening session 2), and when gender was known/provided (listening session 3). A 2×2 (gender $\times$ listening condition) mixed-model analysis of variance was used to determine the influence of known gender on ratings of ACC and NAT, with gender as the between-subjects factor, and listening condition as the within-subjects factor. Significant interactions were interpreted using the least square differences (LSD) procedure for determining simple effects among cell means ($\alpha = .05$).

Reliability for gender ratings, ACC, and NAT

Intrarater reliability was calculated using a Pearson correlation coefficient between initial and repeated samples. Intrarater reliability ranged from $r = .52$ to $r = 1.00$ (mean $r = .89$) for gender ratings with ACC, and from $r = .76$ to $r = 1.00$ (mean $r = .93$) for gender ratings with NAT. Intrarater reliability values ranged from $r = .47$ to $r = .99$ (mean $r = .70$) for ACC when gender was not provided (listening session 2), and from $r = .64$ to $r = .99$ (mean $r = .88$) when gender was provided (listening session 3). Values for NAT ranged from $r = .47$ to $r = .98$ (mean $r = .70$) in listening session 2, and from $r = .65$ to $r = .99$ (mean $r = .84$) in listening session 3.

Interrater reliability was calculated using ICC values. The ICC coefficients for gender ratings were .974 and .979 when they coincided with ACC and NAT ratings, respectively. The ICC coefficient for ACC was .930 when gender was not provided (listening session 2) and .969 when gender was provided (listening session 3). The ICC coefficients for NAT were .939 and .967 for listening sessions 2 and 3, respectively.

Acoustic measures and relationships with perceptual ratings

Pitch-related measures including F_0, SD F_0, min F_0, and max F_0 were calculated using the CSL. If periodic pulses could not be visually identified in a narrowband spectrogram, measures were not calculated.

Duration-related measures included: (1) total duration of the sentence, (2) number of pauses, (3) duration of pauses, and (4) syllables per second. Spectral parameters included SPTILT, TILT-V, LP-SNR, PA, and SFR. Stepwise linear regression analyses were used to determine how well all the acoustic parameters (pitch-related, duration-related, and spectral) predicted the ACC and NAT ratings for male and female TE speakers when gender was unknown (listening session 2) and known (listening session 3).

RESULTS

Paired comparison

Each TE speaker's preference score was based on 220 judgments (11 repetitions $\times$ 20 raters) collapsed across presentation order. Mean preference scores were calculated and ranked from most to least preferred (Figure 1 and Table 2).

Results of the paired comparison task provided a hierarchical ordering of speakers. Mean preference scores ranged from 18.5 for the most preferred speaker (eg, speaker #14) to 40.5 for least preferred speaker (eg, speaker #1). Mean preference scores were not significantly different across male or female speaker groups ($P > 0.05$). The preference rankings were only moderately consistent with the initial selection of the male and female speakers' made relative to overall speech severity[22] (Table 2).

Gender identification, NAT, and ACC ratings

All male speakers were identified as male (6/6). Although two female speakers were identified as female, four of the six females were identified as male. Mean gender ratings are provided in Table 3.

Group means of ACC and NAT ratings when gender was unknown (listening session 2) versus known (listening session 3) are summarized in Table 3. No significant main effects were found for ACC for listening condition, or for the main effect of speaker gender. A significant interaction was found between gender and listening condition for ACC ($F(1,10) = 7.324$, $P < 0.05$). Females were judged less acceptable when gender was known ($M = 16.8$, SD = 11.6) than when gender was unknown ($M = 28.3$, SD = 11.7) (using LSD = 4.398). The ACC of male speakers did not change with listening condition (Table 3).

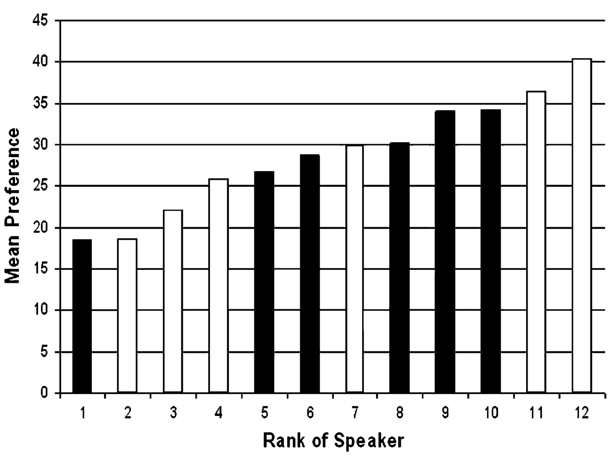

FIGURE 1. Mean preference judgments ranked from most preferred to least preferred. Male speakers are indicated in *filled bars* (black) and female speakers in *unfilled bars* (white).

TABLE 2. *Mean Preference Scores and Rankings of speakers (Out of 12)*

Speaker	Gender	Mean Preference	Preference Rank	Severity Rank
1	F	40.5	12	12
6	F	36.5	11	9
13	F	18.6	2	8
16	F	25.9	4	6
23	F	29.9	7	2
24	F	22.1	3	1
Average		28.7		
2	M	28.6	6	11
4	M	26.7	5	10
14	M	18.5	1	7
18	M	33.9	9	5
19	M	34.1	10	4
21	M	30.1	8	3
Average		28.9		

Note: a smaller score indicates that the speaker was more strongly preferred.

No significant main effects were found for listener judgments of NAT. However, a significant interaction was found between gender and listening condition for NAT ($F(1,10) = 5.382$, $P < 0.05$). Males were judged more natural when gender was known (male $M = 30.5$, SD = 14.4) than when gender was unknown ($M = 24.9$, SD = 10.5) (using LSD = 2.260). In contrast, females were judged to be less natural when gender was known ($M = 16.8$, SD = 12.0) than when gender was unknown ($M = 20.2$, SD = 9.3) (Table 3).

Acoustic measures and relationship with perceptual measures

The objective measures for the speakers are included in Tables 4 and 5. Averages are calculated for each of the female and male speaker groups. In one instance (speaker #2), there was a striking lack of periodicity (gross irregularities in period duration, but also in peak amplitude), even within the vowels of the sentence. The corresponding spectrograms revealed only one or two short-term detectable harmonics. Consequently, pitch-related measures were not calculated for this speaker (Table 5).

TABLE 3. *Speaker Gender, Gender Ratings, ACC and NAT Ratings for Male and Female TE Speakers when Gender was Unknown (Session 2) Versus Known (Session 3)*

		Gender Rating		ACC		NAT	
Speaker	Gender	ACC	NAT	Unknown	Known	Unknown	Known
1	F	25.7	25.0	10.8	5.5	5.8	6.6
6	F	54.6	53.9	21.6	24.6	18.2	29.5
13	F	41.9	43.2	43.0	36.4	34.1	34.5
16	F	20.3	22.9	31.6	14.1	24.8	12.8
23	F	21.2	20.8	25.3	9.1	17.8	8.5
24	F	14.4	13.6	37.9	11.3	20.7	9.0
Average		29.7	29.9	28.3 (11.7)	16.8 (11.6)	20.2 (9.3)	16.8 (12.0)
2	M	4.3	1.0	32.3	42.4	29.4	37.5
4	M	1.4	1.7	36.1	42.7	36.3	43.1
14	M	3.9	1.9	47.9	50.5	38.8	44.4
18	M	12.2	11.5	21.6	24.3	16.0	22.2
19	M	27.5	25.7	23.4	14.0	14.0	15.1
21	M	10.1	7.9	21.4	22.5	15.3	21.2
Average		9.9	8.3	30.4 (11.3)	32.7 (12.6)	24.9 (10.5)	30.5 (14.4)

Note: averages and (standard deviations) are provided across all listener judgments for female and male speakers.
Abbreviations: M, male; F, female. For gender ratings, 1, "definitely male"; 60, "definitely female."

To predict the relationship between the acoustic measures (independent measures) and ACC and NAT judgments (dependent measures) for both male and female speakers, eight separate stepwise linear regression analyses were performed. First, four analyses were performed to predict listener judgments when gender was unknown. For ACC judgments of female speakers, duration of the sentence accounted for 68.5% of the variance ($r^2 = .685$, $F(1,4) = 8.708$, $P < 0.05$), and 86.6% of the NAT judgments ($r^2 = .866$, $F(1,4) = 25.897$, $P < 0.05$). In contrast, ACC and NAT judgments for male speakers when gender was unknown were predicted best with TILT-V. TILT-V predicted 83.3% of ACC judgments ($r^2 = .833$, $F(1,4) = 19.931$, $P < 0.05$) and 76.3% of NAT judgments ($r^2 = .763$, $F(1,4) = 12.881$, $P < 0.05$).

When gender was known, four additional analyses were performed. For ACC of female speakers, 90.1% of the variance was predicted with TILT-V, and an additional 8.4% was predicted with the duration of the sentence, for a total of 98.5% predicted ($r^2 = .985$, $F(2,3) = 98.664$, $P < 0.05$). TILT-V predicted 96.7% of the variance of NAT judgments for females ($r^2 = .967$, $F(1,4) = 118.685$, $P < 0.05$). In contrast, ACC judgments for male speakers when gender

TABLE 4. *Pitch-Related, Duration, and Spectral Measures for Female Speakers*

	F_0 (Hz)	SD F_0 (Hz)	min F_0 (Hz)	max F_0 (Hz)	Duration (s)	# Pauses	Mean Pause Time (s)	Syllable/s	SPTILT	TILT-V	LP-SNR	PA	SFR
1	182.04	69.98	68.12	333.33	6.07	1	0.94	2.47	−20.67	−19.8	−0.13	0.13	−7.42
6	203.95	55.83	69.06	342.47	5.06	0	0	2.96	−7.38	−8.15	−5.28	0.13	−1.63
13	110.37	46.98	61.12	225.23	3.63	0	0	4.14	−5.72	−5.61	−2.07	0.10	−1.42
16	99.25	94.85	61.12	294.12	5.05	2	1.00	2.97	−20.33	−19.77	−0.65	0.11	−4.11
23	106.36	85.78	60.10	333.33	5.22	1	0.39	2.87	−19.49	−19.31	−0.26	0.13	−6.70
24	95.39	65.36	61.12	308.64	5.18	0	0	2.89	−19.39	−18.88	−0.37	0.14	−5.17
Average	132.89	82.01	63.44	306.419	5.04	0.67	0.39	3.05	−15.50	−15.25	−1.46	0.12	−4.41

TABLE 5. *Pitch-Related, Duration, and Spectral Measures for Male Speakers*

	F_0 (Hz)	SD F_0 (Hz)	min F_0 (Hz)	max F_0 (Hz)	Duration (s)	# Pauses	Mean Pause Time (s)	Syllable/s	SPTILT	TILT-V	LP-SNR	PA	SFR
2	N/A	N/A	N/A	N/A	4.67	1	0.50	3.21	−17.95	−17.81	−0.28	0.12	−8.46
4	86.90	70.76	69.06	342.47	5.80	1	1.11	2.59	−9.33	−9.33	−1.69	0.10	−1.48
14	153.18	77.04	61.12	294.12	3.79	0	0	3.96	−21.18	−20.64	−1.83	0.11	−5.17
18	120.51	38.88	74.18	312.50	8.34	0	0	1.80	−21.19	−20.51	−1.33	0.14	−6.90
19	139.46	75.51	61.12	333.33	8.47	4	0.64	1.77	−19.08	−17.91	−1.62	0.16	−8.46
21	95.37	51.23	66.14	260.42	5.21	2	1.11	2.88	−20.79	−20.32	−1.14	0.14	−5.21
Average	119.08	62.68	64.74	298.90	6.05	1.33	0.50	2.70	−18.25	−17.77	−1.31	0.13	−5.42

was known were predicted with two variables. First, 91.6% of the ACC judgments for male speakers was predicted with PA ($r^2 = .916$, $F(1,3) = 32.634$, $P < 0.05$). Next, SFR entered the regression, and predicted an additional 7.6%. Together, 99.2% of the variance was predicted with PA and SFR ($r^2 = .992$, $F(2,2) = 123.804$, $P < 0.05$). NAT judgments of males were also predicted with two variables. PA accounted for 97.2% of the variance for NAT judgments of male speakers ($r^2 = .972$, $F(1,3) = 102.869$, $P < 0.05$). F_0 predicted an additional 2.8%, which resulted in 100% of the variance in NAT judgments of male speakers predicted ($r^2 = 1.000$, $F(2,2) = 8809.260$, $P < 0.05$).

DISCUSSION

The purpose of this study focused on issues related to gender identification and the consequences of gender identification on listener judgments of TE speakers. In designing the present investigation, we actively sought to control several variables that we believed might influence perceptual assessments. First, we attempted to equate speaker performance by attempting to match male and female participants relative to perceived speech severity. Second, direct efforts were made to evaluate listener performance in the context of knowledge related to the gender of speakers. Additionally, listener judgments of ACC and NAT of TE speakers were obtained in an effort to determine relationships between such ratings and speaker gender. Finally, acoustic analyses were performed to help identify objective measures that could be used to determine the ACC and NAT of both male and female speakers. Collectively, the conduct of this investigation sought further information concerning the influence of speaker gender on listener evaluation of TE speech. Based on data gathered, the present findings provide initial insights into a variety of phenomena that would appear to underlie judgments related to gender identification for TE speakers.

Speaker preference

Information related to judgments of speaker preference obtained from the present study is valuable at several levels. First, while the mean preference scores for male and female TE speakers were not found to differ significantly, the procedure used to determine preference may have merit in that listeners were able to sort speakers relative to this global perceptual construct; this paradigm has been used in recent studies.[37,38] In the present study, listeners provided a hierarchical ranking of speakers through ratings of preference. Our results suggest that as a descriptive metric, the perceptual construct underlying ‘‘preference’’ has face validity as a multidimensional feature applied to the assessment of TE signals. Interestingly, the preference judgments yielded a somewhat different ranking of speakers than the overall speech severity judgments from 15 naïve listeners in the initial study.[22] This observation suggests that a listener's preference for particular speakers evolves from perceptual weighting of more than solely the degree of overall speech ‘‘severity.’’ However, the fact that listeners did not inherently prefer one group of speakers (male or female) does suggest that our efforts to control aspects of speech severity across male and female groups were successful and that this factor may have been reduced as a potential confound in other phases of the study.

Although judgments made on the basis of ''preference'' as an overarching perceptual phenomenon may be difficult to rectify from the standpoint of consistently distinguishing male from female TE speakers, the present data suggest that listeners are indeed sensitive to the composite voice/speech signal. That is, despite clear deviance from the ''normal'' perceptual template of the human voice, listeners appear to have the capacity to distinguish TE voices along some type of psychophysical continuum. This finding verifies that listeners may actively modify their internal referent for perceptual judgments of nonnormal speaker groups. This observation has considerable importance relative to how listeners may adjust their intrinsic capacity to evaluate voices that fall well outside of normal expectation. This apparent ability to adjust one's standard in the presence of vocal deviance may be found to correlate with other judgments of communication efficiency and effectiveness within social situations.

Gender identification

The finding that all six male TE speakers were correctly identified as men was not unexpected. We believe that some sets of collective perceptual attributes, whether they emerge from the frequency, amplitude, or temporal domains (or combinations thereof), are perceptually salient to listeners. For example, our male TE speakers had normal to below normal F_0 (as compared to adult male laryngeal speakers), and some degree of vocal noise as indicated by spectral measures. These changes appeared to help listeners perceive male TE speakers as males and are consistent with what is known about correlates of gender in laryngeal speakers.[39]

The somewhat equivocal findings for our group of female TE speakers were consistent with those found by Weinberg and Bennett,[23] who reported that gender identification of male esophageal speakers was more accurate than that for female speakers. However, the present results were not consistent with those reported by Searl and Small.[24] They found that the gender of female TE speakers was identified with equal accuracy as male speakers (about 90%). Although conclusions of gender identification from the present study are difficult to generalize, they have important implications. Whereas most studies have included ''excellent'' or ''superior'' TE speakers, speakers in the present study varied across the speech severity spectrum.[22] Therefore, the results that listeners had difficulty identifying most female TE speakers as women might have more face validity than results from studies which only included superior speakers.[24] These findings clearly require further verification with a larger number of samples. Results gathered from a larger group of speakers would appear to have direct implications for pre- and postlaryngectomy and TE puncture counseling regarding gender identification in instances where the communication partner does not have access to visual or other contextual cues (eg, telephone conversations, instances where communication partners are not facing one another).[24]

The current results also raise additional questions about the two participants who were correctly identified as women, namely, what causes them to be correctly identified? The reciprocal question is also equally germane to those who were misidentified. It would appear that a single, clearly identified variable is rather elusive from the perspective of accurate gender identification. It does not appear unreasonable to suggest that listeners use multiple cues inherent to any given voice signal in their efforts to evaluate and/or categorize it through perceptual processes. For example, studies examining transgender speakers appear to support the notion that subjects who are identified as female exhibit higher mean speaking F_0 and a higher upper limit of speaking F_0.[40] This might help explain the present results for speaker #6 who was clearly identified as female, and who also had an F_0 (203.95 Hz) that was within expected limits for an adult female laryngeal speaker. This explanation does not, however, hold true for speaker #13 who was also identified as female, but had an F_0 of 110.35 Hz. This F_0 value is within the range of a normal adult male laryngeal speaker and is consistent with ''typical'' findings of F_0 for female TE speakers.[8] Although F_0 was low for speaker #13, it is notable that her speaking rate was the closest to normal laryngeal values among all of the speakers. Thus, some combination of rate and F_0 may have aided listeners in identifying this speaker as female. These interpretations are consistent with suggestions provided by both Lanpher[13] and Shanks,[14] who noted that increasing pitch and rate may help a laryngectomized woman sound more feminine.

The very nature of the alaryngeal speaker's vocal tract also may provide a transfer function that assists identification of that signal as that of a man or a woman.[39] Sisty and Weinberg[41] found that female esophageal speakers, like female laryngeal speakers, produce mean formant frequencies that are slightly higher than their male counterparts. Cervera and colleagues[42] have found similar results in TE speaker signals; regardless of gender, TE speakers produce vowels with higher formant frequencies than laryngeal speakers. Sisty and Weinberg[41] suggested that the shift in formant frequencies was due to a reduction in vocal tract effective length postlaryngectomy. Because this shift is systematic, the pattern of spectral changes may be sufficient to serve as a listener cue for gender in some of the current TE speakers. That is, small shifts in frequency for female speakers may, in some instances, assist the listener in distinguishing one's gender; perceptual identification may then be processed in a somewhat unconscious manner so that penalty is applied more commonly to the "noisier" signal produced by female TE speakers. Findings generated from ratings of ACC and NAT in the context of *a priori* gender identification also may shed light on this concern.

ACC and NAT ratings

Speech ACC and NAT are multidimensional constructs that have been used to evaluate alaryngeal speech previously.[5,15,20,21,27,43] Judgments of ACC hold the potential that factors intrinsic to the listener are weighed in the perceptual process. That is, listeners are asked to identify how "acceptable" they find a given signal relative to their own status as a listener. This definition suggests that personal variables may enter into judgments and that such judgments may reflect one's ability to truly "accept" deviation from normal. In this regard, it would seem that knowledge of gender might be a critical factor in determining a speaker's ACC based on listener judgments. This interpretation was supported by the present study in that knowledge of gender substantially penalized only female speakers. Thus, listeners may in fact have one template for men and another for women, not only for unidimensional parameters such as pitch, but also more importantly for composite features that demand more global evaluations and classification of the speech signal.

One concern that may be raised in the context of the current data pertains to how uniformly listeners interpreted the definition provided for the dimension of ACC and NAT. Although the definitions used are longstanding in the literature, the general nature of these definitions does provide the potential for variability in how they are ultimately interpreted and applied by listeners. For example, each of the specific features inherent to the definition of ACC (ie, pitch, understandability, etc) may not be weighted equally in the global evaluation of this dimension by a given listener. Similarly, for the definition of NAT, some of the component areas included in this definition (ie, rhythm, intonation, etc) may not necessarily have perceptual salience in a strict fashion. In this instance, some listeners might not understand what constitutes "intonation" and may then move toward a more global assessment of NAT as a feature that might more closely reflect its comparison to a "natural" (ie, normal sounding) vocal signal. This possibility was indeed validated through the high correlations exhibited between ACC and NAT judgments regardless of whether gender was known ($r = .91$ when gender was unknown; $r = .98$ when gender was known). These correlations may indicate that similar perceptual constructs were addressed at some level.

Once knowledge of gender was introduced into the experimental scheme, an alteration in the listener's evaluative template may have occurred. This suggestion also may account for our observation that male speakers were judged more favorably for NAT when gender was known. For ACC, knowledge of gender resulted in dramatic differences in ratings for females when compared to males; ratings for males remained relatively constant regardless of whether gender was known. For NAT, however, ratings for males improved and those for females decreased when gender was known. This finding demonstrates that some signal feature(s) differentially influence(s) judgments of NAT. That is, perceptual ratings do not change in only one direction for NAT when gender is known—ratings for men improve and those for women decrease. When compared to ACC ratings, it can be seen that relative stability was noted for men, but women were penalized considerably. For this reason, further explorations directed toward familiarizing the listener to definitional boundaries, or educating the listener relative to a specific dimension would seem appropriate.

Because ACC may be influenced by an individual listener's personal level of "tolerance" and as suggested earlier, by expectation in the gender-known condition, such ratings may reflect even greater differences for females. But the presence of these differences may exist despite knowledge of a given speaker's gender. For example, it is interesting to note that regardless of whether ACC or NAT was being evaluated, the mean scaled scores were lower for women (Table 3). These results may indicate that in instances where a uniform standard is used (ie, in the gender unknown condition), some perceptual distinction does exist between men and women. Although overlap in these scaled scores did exist for both dimensions, ACC was reduced for five of six women, whereas only one male exhibited a reduction in ACC when gender was known to the listener. A similar trend existed for NAT, in that three of the women's scores were reduced (two were only marginally better) when gender was known; all six males were judged more favorably when gender was known. Though both perceptual dimensions seem to permit listeners to make distinctions, the underlying reason for these changes in judgment remains elusive. Acoustic cues associated with these judgments may aid in the interpretation of these results.

Acoustic measures and perceptual ratings

While no single link between acoustic and perceptual ratings was identified in the present study, some findings deserve comment. First, for female TE speakers, data analysis found that the duration of the sentence accounted for a relatively large percentage of variance (69% and 87%) for ACC and NAT judgments, respectively, when gender was unknown. Interestingly, F_0 was not found to correlate strongly with listener judgments of females. Rather, aspects of "timing" and flexibility of the speech signal were factors that appeared to provide some link to judgments of female TE speakers. These results are consistent with those found in the dysarthria and stuttering literature, in that speech NAT has consistently been associated with speaking rate.

On the other hand, judgments of ACC and NAT for male speakers were based on spectral features (83% and 76% predicted, respectively) when gender was unknown. Thus, it would appear reasonable to suggest that something about the female TE signal appears to lead the listener to judge female signals in a different fashion from those of their male counterparts. In this regard, we believe that a variety of temporal variables within a given speech structure (ie, sentence stimuli or longer samples of ongoing speech) and how these variables interact may provide a potential source of increasing a listener's perception of the NAT and ACC of female TE speech samples. This interpretation may provide support for previously recommended therapy approaches to increase the femininity of female esophageal speakers.[14]

Results appear to indicate that when gender was known to listeners, judgments of females changed from being associated with temporally based features to those based on spectral cues. In this condition (gender known), listeners appeared to use similar cues for making ACC and NAT judgments, regardless of speaker gender. This interpretation was supported both by relationships with acoustic cues (eg, use of similar acoustic cues such as PA and spectral tilt) and the strong correlation between judgments of ACC and NAT when gender was known ($r = .98$). Hence, spectral dimensions of the TE signal appear to influence perceptual judgments most favorably for male TE speakers. These findings are consistent with those found by Trudeau and Qi,[8] who found that the greatest difference between female TE voices and male TE voices was related to perturbation measures.

Results from the acoustic analysis, in combination with the known moderate strength of relationships between listener judgments when gender was unknown or known ($r = .67$ for ACC; $r = .81$ for NAT), appear to support the interpretation that listeners in the "gender-known" condition may shift their perceptual strategy in this condition. That is, when gender was known, the point of reference was that of the normal laryngeal signal. In particular, listeners may have reinterpreted both dimensions such that cues used for making NAT judgments appeared to be similar to those used for ACC when gender was known. In this condition, listeners seemed to focus on signal ''noise'' as indicated by the number of spectral cues that were associated with these judgments. These features seemed to override other cues such as temporally based measures or pitch-related factors that may have served to differentiate male and female speakers, or to differentiate judgments of ACC and NAT. Further investigation of these dimensions using experienced listeners, who might not be as distracted by the ''noisy'' TE signal, might be warranted.

CONCLUSION

This study evaluated a variety of factors related to gender identification in TE speakers. Based on the present data, it appears that naïve listeners did not exhibit a clear preference for either male or female TE speakers. These data support our effort to control the variable of overall speech severity for the two groups of speakers. Listeners also had difficulty identifying the gender of most female speakers. When listeners were provided with information related to the gender of the speakers, they differentially penalized female TE speakers as significantly less acceptable and natural than their male counterparts. While the reasons for this difficulty are unknown, factors specific to the temporal domain appear to contribute substantially to the perceptual process for both judgments of ACC and NAT in female speakers. In contrast, spectral factors appear to primarily influence judgments of ACC and NAT in TE speakers, regardless of gender, when gender is known. Collective evaluation of both acoustic and perceptual judgments should continue to be investigated to further understand variables that may underlie and influence the perception of speaker gender by naïve listeners. These differences are critical to understand if we are to promote successful rehabilitation of both men and women who undergo total laryngectomy.

REFERENCES

1. Singer MI, Blom ED. An endoscopic technique for restoration of voice after laryngectomy. *Ann Otol Rhinol Laryngol*. 1980;89:529–533.
2. Iversen-Thoburn SK, Hayden PA. Alaryngeal speech utilization: a survey. *J Med Speech Lang Pathol*. 2000;8:85–99.
3. Doyle PC. *Foundations of Voice and Speech Rehabilitation Following Laryngeal Cancer*. San Diego, CA: Singular Publishing Group; 1994.
4. Bellandese MH, Lerman JW, Gilbert HR. An acoustic analysis of excellent female esophageal, tracheoesophageal, and laryngeal speakers. *J Speech Lang Hear Res*. 2001;44:1315–1320.
5. Finizia C, Dotevall H, Lundström E, Lindström J. Acoustic and perceptual evaluation of voice and speech quality. *Arch Otolaryngol Head Neck Surg*. 1999;125:157–163.
6. Hillman RE, Walsh MJ, Wolf GT, Fisher SG, Hong WK. Functional outcomes following treatment for advanced laryngeal cancer. *Ann Otol Rhinol Laryngol*. 1998;107:2–27.
7. Robbins J, Fisher HB, Blom EC, Singer MI. A comparative acoustic study of normal, esophageal, and tracheoesophageal speech production. *J Speech Hear Disord*. 1984; 49:202–210.
8. Trudeau MD, Qi YY. Acoustic characteristics of female tracheoesophageal speech. *J Speech Hear Disord*. 1990; 55:244–250.
9. Williams SE, Watson JB. Differences in speaking proficiencies in three laryngectomee groups. *Arch Otolaryngol*. 1985;111:216–219.
10. Pindzola RH, Cain BH. Duration and frequency characteristics of tracheoesophageal speech. *Ann Otol Rhinol Laryngol*. 1989;98:960–964.
11. Van As CJ, Hilgers FJM, Verdonck-de Leeuw IM, Koopmans-van Beinum FJ. Acoustical analysis and perceptual evaluation of tracheoesophageal prosthetic voice. *J Voice*. 1998;12:239–248.
12. Finizia C, Hammerlid E, Westin T, Lindström J. Quality of life and voice in patients with laryngeal carcinoma: a post-treatment comparison of laryngectomy (salvage surgery) versus radiotherapy. *Laryngoscope*. 1998;108:1566–1573.
13. Lanpher A. Making the esophageal voice feminine. In: Lauder E, ed. *Self Help for the Laryngectomee*. San Antonio, TX: Lauder Enterprises; 1997:71–75.
14. Shanks J. Development of feminine voice and refinement of esophageal voice. In: Keith RL, Darley FL, eds. *Laryngectomee Rehabilitation*. 2nd ed. San Diego, CA: College Hill; 1986:269–276.
15. Doyle PC, Eadie TL. The perceptual nature of alaryngeal voice and speech. In: Doyle PC, Keith RL, eds. *Contemporary Considerations in the Treatment and Rehabilitation of Head and Neck Cancer: Voice, Speech, and Swallowing*. Austin, TX: Pro-Ed; 2005:113–140.
16. Shipp T. Frequency, duration, and perceptual measures in relation to judgments of alaryngeal speech acceptability. *J Speech Hear Res*. 1967;10:417–427.

17. Weinberg B, Bennett S. Selected acoustic characteristics of esophageal speech produced by female laryngectomees. *J Speech Hear Res*. 1972;15:211–216.
18. Snidecor JC, Curry ET. Temporal and pitch aspects of superior esophageal speech. *Ann Otol Rhinol Laryngol*. 1959;68:1–14.
19. Van Riper C. *Speech Correction: Principles and Methods*. Englewood Cliffs, NJ: Prentice Hall; 1978.
20. Bennett S, Weinberg B. Acceptability ratings of normal, esophageal, and artificial larynx speech. *J Speech Hear Res*. 1973;16:608–615.
21. Pindzola RH, Cain BH. Acceptability ratings of tracheoesophageal speech. *Laryngoscope*. 1988;98:394–397.
22. Eadie TL, Doyle PC. Auditory-perceptual scaling and quality of life in tracheoesophageal speakers. *Laryngoscope*. 2004;114:753–759.
23. Weinberg B, Bennett S. A study of talker sex recognition of esophageal voices. *J Speech Hear Res*. 1971;14:391–395.
24. Searl JP, Small LH. Gender and masculinity-femininity ratings of tracheoesophageal speech. *J Commun Disord*. 2002;35:407–420.
25. Kreiman J, Gerratt BR, Kempster GB, Erman A, Berke GS. Perceptual evaluation of voice quality: review, tutorial, and a framework for research. *J Speech Hear Res*. 1993;36:21–40.
26. Berlin CI. Clinical measurement of esophageal speech: III. Performance of non-biased groups. *J Speech Hear Dis*. 1965;30:174–183.
27. Filter MD, Hyman M. Relationship of acoustic parameters and perceptual ratings of esophageal speech. *Percept Mot Skills*. 1975;40:63–68.
28. Most T, Tobin Y, Mimran RC. Acoustic and perceptual characteristics of esophageal and tracheoesophageal speech production. *J Commun Disord*. 2000;33:165–180.
29. Snidecor JC. Some scientific foundations for voice restoration. *Laryngoscope*. 1975;85:640–648.
30. American Cancer Society. *Cancer Facts and Figures—2003*. Atlanta, GA: American Cancer Society; 2003.
31. Fairbanks G. *Voice and Articulation Drillbook*. 2nd ed. New York, NY: Harper; 1960.
32. Avaaz Innovations Inc. *Experiment Controller and Generator for Windows (Ecos/Win)* [Computer Program]. London, Ontario, Canada: Avaaz Innovations Inc.; 1998.
33. Yorkston KM, Beukelman DR, Bell KR. *Clinical Management of Dysarthric Speakers*. Boston, MA: College-Hill Press; 1988.
34. Kay Elemetrics Corporation. *Computerized Speech Lab (CSL), Model 5400*. Lincoln Park, NJ: kay Elemetrics Corporation; 2000.
35. Avaaz Innovations Inc. *Interactive Voice Analysis System (IVANS), Version 1.0* [Computer Program]. London, Ontario, Canada: Avaaz Innovations Inc.; 1998.
36. Parsa V, Jamieson DG. Acoustic discrimination of pathological voice: sustained vowels versus continuous speech. *J Speech Lang Hear Res*. 2001;44:327–339.
37. Meltzner GS, Hillman RE. Impact of aberrant acoustic properties on the perception of sound quality in electrolarynx speech. *J Speech Lang Hear Res*. 2005;48:766–779.
38. Graville DJ, Cohen JI, Rau MT. Acoustic parameters that identify listeners' perceptions of good near-total laryngectomy voice. *J Med Speech Lang Pathol*. 2004;12:107–116.
39. Mullenix JW, Johnson KA, Topcu-Durgun M, Farnsworth LM. The perceptual representation of voice gender. *J Acoust Soc Am*. 1995;98:3080–3095.
40. Gelfer MP, Schofield KJ. Comparison of acoustic and perceptual measures of voice in male-to-female transsexuals perceived as female versus those perceived as male. *J Voice*. 2000;14:22–33.
41. Sisty NL, Weinberg B. Formant frequency characteristics of esophageal speech. *J Speech Hear Res*. 1972;15:439–448.
42. Cervera T, Miralles JL, Gonzalez-Alvarez J. Acoustical analysis of Spanish vowels produced by laryngectomized subjects. *J Speech Lang Hear Res*. 2001;44:988–996.
43. Tardy-Mitzell S, Andrews M, Bowman SA. Acceptability and intelligibility of tracheoesophageal speech. *Arch Otolaryngol*. 1985;111:212–215.