Impact of Aberrant Acoustic Properties on the Perception of Sound Quality in Electrolarynx Speech

Geoffrey S. Meltzner
Massachusetts Institute of Technology,
Cambridge

Robert E. Hillman
Massachusetts General Hospital and
Harvard Medical School, Boston

A large percentage of patients who have undergone laryngectomy to treat advanced laryngeal cancer rely on an electrolarynx (EL) to communicate verbally. Although serviceable, EL speech is plagued by shortcomings in both sound quality and intelligibility. This study sought to better quantify the relative contributions of previously identified acoustic abnormalities to the perception of degraded quality in EL speech. Ten normal listeners evaluated the sound quality of EL speech tokens that had been acoustically enhanced by (a) increased low-frequency energy, (b) EL-noise reduction, and (c) fundamental frequency variation to mimic normal pitch intonation in relation to nonenhanced EL speech, normal speech, and normal monotonous speech (fundamental frequency variation removed). In comparing all possible combinations of token pairs, listeners were asked to identify which one of each pair sounded most like normal natural speech, and then to rate on a visual analog scale how different the chosen token was from normal speech. The results indicate that although EL speech can be most improved by removing the EL noise and providing proper pitch information, the resulting quality is still well below that of normal natural speech or even that of monotonous natural speech. This suggests that, in addition to the widely acknowledged acoustic abnormalities examined in this investigation, there are other attributes that contribute significantly to the unnatural quality of EL speech. Such additional factors need to be clearly identified and remedied before EL speech can be made to more closely approximate the sound quality of normal natural speech.

KEY WORDS: electrolarynx speech, alaryngeal speech, laryngectomy speech rehabilitation, speech quality, artificial larynx speech

The electrolarynx (EL) is an electromechanical device that acoustically excites the vocal tract via neck or mouth placement. The EL is used primarily by individuals who have had the larynx removed due to cancer (patients with laryngectomy) and need an alternative voicing source in order to speak. There is a wide variation in the reported usage of the EL among alaryngeal speakers. Some studies have reported that a minority of patients with laryngectomy use EL speech as their primary means of communication, with estimates of EL use ranging from 11% to 34% (Diedrich & Youngstrom, 1977; Gates, Ryan, Cantu, & Hearne, 1982; Gates et al., 1982; King, Fowlks, & Peirson, 1968; Kommers & Sullivan, 1979; Richardson & Bourque, 1985; Webster & Duguay, 1990). Conversely, other studies have shown that a majority of patients with laryngectomy use some type of EL to

1092-4388/05/4804-0766

communicate, with estimates of EL use ranging from 50% to 66% (Gray & Konrad, 1976; Hillman, Walsh, Wolf, Fisher, & Hong, 1998; Morris, Smith, Van Demark, & Maves, 1992). Even though the prevalence of EL speech may vary among specific subpopulations of individuals with laryngectomy, it is clear that EL devices continue to represent an important option for speech rehabilitation. Even in cases where esophageal or tracheo-esophageal (TEP) speech is ultimately developed, EL devices may serve early on to provide a viable and relatively rapid method of postlaryngectomy oral communication (Hillman et al., 1998). It is also not uncommon for the EL device to continue to serve as a reliable backup in instances where individuals experience difficulties with use of esophageal or TEP speech.

Even though the electrolarynx generally provides a serviceable means of communication, the resulting speech has several shortcomings in terms of both intelligibility and speech quality. *Intelligibility* refers to the recognizability of a speech stimulus, whereas *speech quality* concerns the acceptability of the speech to listeners. Although the two concepts are related, they are treated separately in this study. Reduced intelligibility has mainly been attributed to voicing feature confusion for stop consonants, although vowel confusion also plays a role (Weiss, Yeni-Komshian, & Heinz, 1979). Lack of fundamental frequency (pitch) variation, the presence of a competing self-generated EL noise, and an improper source spectrum have all been cited as important contributors to the poor quality of EL speech (Espy-Wilson, Chari, MacAuslan, Huang, & Walsh, 1998; Qi & Weinberg, 1991, Uemi, Ifukube, Takahashi, & Matsushima, 1994), which has been described as mechanical and nonhuman sounding.

The basics of current EL technology were introduced more than 40 years ago (Barney, Haworth, & Dunn, 1959), but until relatively recently there has been little effort to remedy the primary deficits associated with EL speech. Qi and Weinberg (1991) attempted to improve the quality of EL speech by enhancing its low-frequency content. They developed an optimal second-order low-pass filter to compensate for the "low-frequency deficit" in EL speech and found that the resulting speech was preferred over the original EL speech.

Cole, Sridharan, Moody, and Geva (1997) demonstrated that a combination of noise reduction algorithms (spectral subtraction and root cepstral subtraction) originally developed for the removal of noise corruption in speech signals could be used to effectively remove the ambient noise generated by the EL device (referred to as *EL noise* in the remainder of this article) from audio recordings of EL speakers. Nevertheless, the perceptual improvement afforded by this noise reduction algorithm was modest at best. The improved speech produced a mean opinion score of 2.8 (on a 1 to 5 scale) while the unaltered EL speech produced a score of 2.3. Espy-Wilson et al. (1998) used a somewhat different approach to remove the EL noise. They simultaneously recorded the output at both the lips and at the EL, and then used both signals in an adaptive filtering algorithm to remove the directly radiated EL noise. Perceptual experiments revealed that listeners generally preferred the filtered speech as compared to the unprocessed speech, but at the cost of reduced intelligibility of word-initial nasals.

Uemi et al. (1994) designed a device that used air pressure measurements obtained from a resistive component placed over the stoma to control the fundamental frequency of an EL. In an even more ambitious approach, Ma, Demirel, Espy-Wilson, and MacAuslan (1999) used cepstral analysis of speech to replace the EL excitation signal with a normal speech excitation signal, while keeping the vocal tract information constant. Not only did the normal excitation signal contain the proper frequency content (i.e., no low-frequency deficit), but it also contained a natural pitch contour to help eliminate the monotonous quality of EL speech.

The success of these studies indicates that EL users could gain some benefit from an EL communication system that improves the quality of the speech in one of these ways. However, each of these enhancements has only been tried in isolation, and some are more difficult to implement than others. Thus, knowing the relative contribution that these different enhancements make (both alone and in combination) to improve the perceived quality of EL speech is critical in determining which approaches should be given priority in future attempts to actually implement such enhancements in a device that patients can use. Moreover, formally assessing how closely the perceived quality of the best enhanced EL speech approximates normal natural speech would indicate the limits of previous proposed enhancement approaches and serve to estimate how much more potential there is for further improving EL speech. The goals of this investigation were to better quantify the sources and perceptual impact of abnormal acoustic properties typically found in EL speech by (a) quantifying the relative contribution that acoustic enhancements make, both individually and in combination, to improving the perceived quality of EL speech and (b) determine how closely the best enhanced EL speech approximates normal-natural speech quality. This investigation was part of a large ongoing project aimed at developing an improved EL communication system (Meltzner et al., 2005).

Method

Recording Procedures

Two normal (i.e., nonlaryngectomized) speakers, 1 male and 1 female, who were very experienced at speaking with an EL, served as participants. The initial step for each participant was to determine the neck location at which the EL produced the clearest sounding speech (patients with laryngectomy refer to this as the "sweet spot"). This site was then marked so that the EL could be placed at this location for all recorded tasks. All EL recordings were made with a neck-placed Servox Inton EL (Siemens Corp.). The speakers were instructed to breath hold and maintain a closed glottis while talking with the Servox to approximate the anatomical condition of patients with laryngectomy in which the lower airway is disconnected from the upper airway.

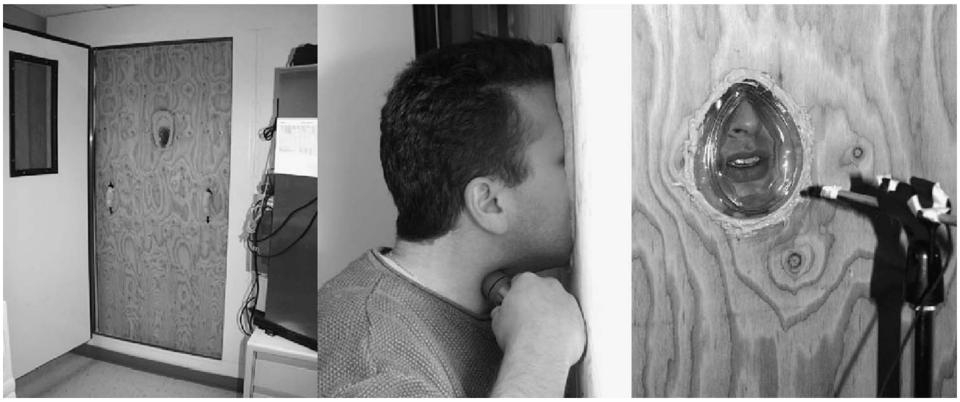
Participants were recorded under two conditions: (a) seated in a sound isolated booth and (b) standing outside of a sound isolated booth with his or her face sealed in the port of a specially constructed door attached to the booth (see Figure 1). In both conditions the microphone was located in the sound isolated booth.

The first condition was used to obtain free-field recordings of the EL noise while the speakers held the EL to their necks and kept their mouths closed. The second condition was designed to essentially eliminate the EL noise associated with neck placement of the device from the audio recording of the speech. This second recording condition reduced the EL noise by 32.4 ± 11.5 dB on average across frequency and speakers. All recordings were made with a Sennheiser (Model K3-U) microphone placed 15 cm from the lips. The participants were asked to say two sentences: (1) "We were away a year ago when I had no money" and (2) "She tried the cap and fleece so she could pet the puck." The lengths of both sentences were chosen so that they could be easily spoken in a single breath (Crystal & House, 1982; Mitchell, Hoit, & Watson, 1996). This was done to prevent the speakers from inserting pauses in the speech, which could provide listeners with additional cues to distinguish between the different types of speech tokens. The different phonemic make up of the two sentences allowed an examination of whether the perceived quality of the different types of speech tokens was impacted by an all-voiced (first sentence) versus a voiced–voiceless (second sentence) environment.

The microphone signals were low-pass filtered at 20 kHz by a 4 pole Bessel filter (Axon Instruments Cyberamp) and then directly digitized at 100 kHz to an American Micro Devices Athlon-based PC system via an Axon Instruments Digidata acquisition board and accompanying Axoscope software. The signals were then appropriately low-pass filtered and downsampled to 8 kHz so that they would meet the bandwidth requirements of the mixed excitation linear predictive vocoder, which is described in the next section.

The use of speakers with normal vocal tract anatomy to produce the experimental sentences deviates from the typical use of an EL in a number of ways. Although EL users generally use the device on a regular, if not everyday, basis, this was not the case with experimental talkers. Moreover, the anatomy of these

Figure 1. Three views of the specially constructed door used for the second recording condition. Left: The specially constructed door with the face port sealed in place in the doorway of the acoustic chamber. Middle: A speaker using the door while speaking with an electrolarynx (EL). Keeping the EL outside of the acoustic chamber reduces the amount of self-noise in the resulting speech. Right: A view from inside the acoustic chamber. The plastic mask seals around the speaker's nose and mouth while a microphone placed inside the booth records the speech.



speakers deviated from that of usual EL speakers (i.e., patients with laryngectomy) in that the larynx and its surrounding structures were intact. Aside from these obvious differences, use of normal laryngeal speakers also meant that there was the potential for subglottal coupling to occur, which is precluded by the laryngectomy procedure that effectively separates the subglottal system from the rest of the vocal tract The lack of coupling affects the acoustics of the vocal tract mainly by narrowing the bandwidths (Meltzner, 2003). Additionally, the laryngectomy procedure involves the truncation/shortening of the supraglottal vocal tract, thereby shifting the formants to higher frequencies (Sisty & Weinberg, 1972).

Nevertheless, the use of normal speakers was necessary for performing this study. The goal of this research was not only to gauge the relative contributions of acoustic abnormalities to the perceived quality of EL speech but also to determine how closely different methods for enhancing EL speech could approximate the quality of normal natural speech. Although comparisons could be made between the EL speech of participants with laryngectomy and the normal speech of participants without laryngectomy, such comparisons could be confounded by the inherent differences in speech acoustic parameters that result from normal between-speaker variations in vocal tract morphology and articulatory dynamics (e.g., speech rate or phoneme durations). Using the same speakers for both the normal and EL speech conditions enabled perceptual judgments to be limited to within-speaker comparisons, thereby greatly reducing the confounding influences of interspeaker variation. In addition, even though the EL sentences still needed to be time aligned with the normal sentences, reducing the amount of required stretching and compression reduced the possible distortion that could occur from this process.

Moreover, it was possible to minimize most of the effects produced by using normal laryngeal speakers. Both participating speakers had extensive previous experience with using an EL, were adept at speaking with the device, and were able to carry on an intelligible conversation without any difficulty. Furthermore, speakers with normal vocal tract anatomy can approximate the anatomy of a patient with laryngectomy and the acoustic effects of that anatomy by maintaining a closed glottis position (House & Stevens, 1958). The only difference that could not be adequately compensated for was the shorter vocal tract lengths of patients with laryngectomy and the associated higher formants. Again, however, eliminating this formant shift was desirable, as it removed a potential confounding perceptual cue that listeners could have used to further distinguish between normal and EL speech.

Generation of Sentence Stimulus Tokens

For each speaker, 10 versions of each recorded sentence were generated: a normal version, a normal version with a fixed/mono fundamental frequency (F_0), raw EL speech, and all possible combinations of three types of EL speech enhancement: (a) boosting of energy below 500 Hz (L), (b) EL noise reduction (N), and (c) F_0 variation to mimic normal pitch intonation, which is designated by a (P) and referred to in the remainder of this article as *pitch*. The notation used to specify each of the 10 different types of sentence tokens is summarized in Table 1.

The boosting of energy below 500 Hz was implemented by processing the sentences through the two-pole low-pass filter specified by Qi and Weinberg (1991):

$$H(z) = \frac{1}{(1 - az^{-1})^2}, \tag{1}$$

where $a = 0.81$.

Speaking through the port in the door tended to slightly restrict articulatory movements of the jaw and lips. Although the perceptual effect of this restriction was small, we decided to use only those versions of sentences (both normal and EL) that were recorded under this condition so as to remove even slight differences in articulation as potential perceptual cues. Thus, to construct stimuli representing unprocessed/raw EL speech, a time-aligned estimate of the EL noise was added to the EL sentences that were recorded through the port of the sound isolated booth. The EL noise estimates were made from the previously described free-field, closed-mouth recordings in the sound isolated booth. The amplitude of the EL noise estimate was scaled such that the resulting speech sounded virtually identical to the unprocessed EL

Table 1. Notation and description of sentence stimuli.

Sentence version	Sentence description
EL-raw	Unprocessed EL speech
EL-L	EL speech with increased low-frequency energy
EL-N	EL speech with noise reduction
EL-P	EL speech with pitch modulation
EL-LN	EL speech with increased low-frequency energy and noise reduction
EL-LP	EL speech with increased low-frequency energy and pitch modulation
EL-NP	EL speech with noise reduction and pitch modulation
EL-LNP	EL speech with all three enhancements
Norm-mono	Monotonous (fixed pitch) normal speech
Normal	Normal natural speech

speech recorded in the sound booth. Time synchronization was performed by aligning the first peak of the EL noise signal with that of the recorded speech. Since the same EL fundamental frequency was used in both recording conditions, the peaks and troughs in the time signals of both the EL noise and the speech lined up perfectly. The assumption in this reconstruction that the EL noise is purely additive has been used in previous studies for the purpose of noise removal (e.g., Meltzner, Kobler, & Hillman, 2003).

The addition of the proper pitch variation to the EL speech involved three steps. First, the normal and EL sentences were time aligned using the pitch-synchronous overlap-add algorithm (Moulines & Charpentier, 1990), so that the phonemes of both sentences had the same onset times and durations. Both sentences were then analyzed using a modified version of a mixed excitation linear predictive (MELP) vocoder (McCree & Barnwell, 1995; U.S. Department of Defense, 1999). The MELP vocoder was chosen for this task because it effectively separates speech into source and filter parameters that are easily manipulable, while producing high quality resynthesized speech. Finally, the F_0 track obtained from the MELP analysis of the normal sentence was used in the MELP synthesis of the EL speech, thus giving the EL sentence the same exact F_0 contour as that of the normal sentence. Because the second sentence contained unvoiced phonemes, there were sections in which no F_0 estimate could be made during MELP analysis. Therefore, before the measured F_0 contour was used in the resynthesis of the EL sentences, the sections of the F_0 contour corresponding to the unvoiced sections were set equal to the last F_0 measurement made prior to the onset of each unvoiced section. As a result, the F_0 was set at a fixed value during what were the unvoiced sections of the normal version of the voiced/voiceless sentence. Moreover, during the resynthesis of the EL versions of this sentence, every frame was set as voiced. The all voiced nature of this sentence is consistent with typical EL speech, because users generally cannot adequately control the EL source well enough to turn it off during voiceless phonemes.

The MELP vocoder was also used to set the constant F_0 of the monotonous EL sentences to the mean F_0 of the normal sentences. This step was taken to remove the potentially confounding influence that differences in the pitches of the stimuli might have on perceptual comparisons. Similarly, the monotonous normal speech token was generated by fixing the F_0 of the whole sentence at the mean F_0. It should be noted that for the female speaker, implementing this step meant that the F_0 would be at a level above what a Servox EL is able to produce, in effect making the EL speech sentences better than they really should be. However, it was decided that removing differences that could act as perceptual cues was more important than keeping the F_0 within the Servox range.

Experimental Procedure

The experimental procedure used the method of paired comparisons (PC) (Torgerson, 1957) and accompanying visual analog scaling (VAS). For each of the four speaker (male or female) and sentence (all voiced or voiced–voiceless) conditions, all combinations of pairs of speech tokens (45) were presented via the data acquisition computer equipped with an Aureal Vortex soundcard and headphones (Sennheiser HD280 Pro) to a group of 10 naïve, normal hearing adult listeners (5 male and 5 female). Before making judgments within each of the four speaker–sentence conditions, all 10 speech tokens for that condition were played to the listeners to familiarize them with the quality of all of the different types of speech tokens that would be judged. Once the experimental procedure began, listeners were required to indicate on a computer response screen which of the two tokens in each pair "sounded more like normal natural speech" (NNS). Once this decision was made, the listener was then asked to use a mouse-controlled VAS to rate how different the chosen token was from NNS. The scale was 10 cm long, with the left end labeled *Not At All Different* and the right end labeled *Very Different*. The distance (in centimeters) from the left end of the scale was used as the rating of the stimulus. The additional VAS procedure provided a way to corroborate the PC-based scaling of the sentence tokens and allowed formal testing for statistically significant differences in the perceived quality of the different sentence tokens. Each complete set of tokens pairs was presented twice in different random orders to assess listener reliability, so that a total of 360 sentences were presented to each listener. The presentation order of the four speaker–sentence conditions was randomized for each listener. The listeners were allowed to adjust the loudness of the sentences to a comfortable level. Once the experiment began, participants were allowed to listen to the normal token (anchor) as often as they wanted during both the PC and VAS components of the experiment, thus providing a consistent and common frame of reference to use in making their judgments.

Data Analysis

Analysis of paired comparison data: Law of comparative judgment. The data collected from the

PC procedure were analyzed using Thurstone's law of comparative judgment (Thurstone, 1927). It is assumed that for each participant a group of stimuli elicits a set of discriminal processes (or perceptions) along a psychological continuum with respect to a certain attribute of the stimuli. However, because human observers tend to be inconsistent, a stimulus will not always elicit the same discriminal process every time it is presented. As such, the most common process is labeled the *modal discriminal process,* whereas the spread of the discriminal process is called the *discriminal dispersion*. If these discriminal processes are modeled as normal random variables, then the modal discriminal processes and the discriminal dispersions are the mean and standard deviation of the random variables, respectively, where the mean is taken to be the scale value on the psychological continuum.

If two stimuli, j and $k,$ are presented to a group of several listeners, and stimulus j is chosen more often to be "greater" (in terms of a certain attribute) than stimulus $k,$ then it can be assumed that the scale value, $S_{j},$ of stimulus $j,$ is greater than the scale value, $S_{k},$ of stimulus k. Furthermore, the proportion of times that stimulus j is chosen over stimulus k is related to the difference between the scale values (i.e., the discriminal difference). This discriminal difference is also a normal random variable with a mean of $S_j - S_k$ and standard deviation of

$$\sigma_{j-k} = \sqrt{\sigma_j^2 + \sigma_k^2 - 2r_{jk}\sigma_k\sigma_j}, \qquad (2)$$

where s_j and s_k are the discriminal dispersions of stimuli j and $k,$ respectively, and r_{jk} is the correlation between the two stimuli. It then follows that the discriminal dispersion between two stimuli can be calculated from

$$S_j - S_k = z_{jk}\sqrt{\sigma_j^2 + \sigma_k^2 - 2r_{jk}\sigma_k\sigma_j}, \qquad (3)$$

where z_{jk} is the normal deviate corresponding to the theoretical proportion by which stimulus j is judged as greater than stimulus k. Because the theoretical values of z_{jk} are not available, they are estimated from the empirical values obtained from the paired comparisons experiment. Equation 3 represents the complete version of Thurstone's law of comparative judgment. It is, unfortunately, impossible to solve Equation 3 because there will always be a larger number of unknowns than observable equations (Torgerson, 1957) and, thus, some simplifying assumptions must be made. Thurstone (1927) discussed several different cases of simplifications; however, the present discussion is restricted to Thurstone's Case V, where it is assumed that the discriminal dispersions are equal and that correlations between stimuli are also equal. This reduces Equation 3 to

$$S_j - S_k = z_{jk}\sqrt{2\sigma^2(1-r)}. \qquad (4)$$

The term $\sqrt{2\sigma^2(1-r)}$ is a scaling constant and can be set equal to 1 without any loss of generality (Edwards, 1957) so that

$$S_j - S_k = z_{jk}. \qquad (5)$$

Hence, the scale value of each stimulus can be found, thus providing not only a ranking of the stimuli but the psychological distance between them on the psychological continuum.

The following procedure is used to generate z_{jk}. The proportion of times stimulus j is judged as greater than stimulus $k,$ p_{jk} is entered into the jth column and kth row of a matrix, $P,$ such as the one shown in Table 2. Because no stimulus is ever presented against itself, the diagonals of the P matrix remain empty. The Z matrix, whose cells contain $z_{jk},$ is found by computing the normal deviates of the entries in the P matrix. The diagonal entries of the Z matrix are set to zero. If the Z matrix is full (i.e., there are no infinite values in any of the entries), then the S_j are easily computed by averaging each column of the Z matrix. However, in many circumstances, one stimulus is always judged to be "better" (or "worse") than another, thereby producing a proportion, $p_{jk},$ of 1 (or 0) and a corresponding infinite z_{jk}. In such cases, simply averaging the columns of the Z matrix is not possible, and another method of estimating the scale values must be used. Kaiser and Serlin (1978) suggested a least squares method to estimate the scale values that was valid as long as the data collected from every stimulus are at least indirectly connected to each other, that is, as long as no stimulus is always judged to be better (or worse) than all the others. When the Z matrix is full, the Kaiser–Serlin method reduces to averaging the columns of the matrix.

TBL2

Unfortunately, because of the nature of the stimuli used in this experiment, the necessary condition for using the Kaiser–Serlin method was, in some instances, violated. Specifically, for some speaker–sentence

Table 2. An example of the *P* matrix.

Stimulus	1	2	…	*n*
1	-	p_{21}	…	p_{n1}
2	p_{12}	—	…	p_{n2}
…	…	…	—	…
n	p_{1n}	p_{2n}	…	—

Table 3. The overall paired comparison scale and visual analog scale values.

Paired Comparison Scale Values			Visual Analog Scale Values			
Speech Type	Rank	Rating	Speech Type	Rank	Rating	s_m
EL-raw	10	0.00	EL-raw	10	8.4	0.20
EL-L	9	0.87	EL-L	9	8.08	0.19
EL-N	8	3.62	EL-LN	8	7.93	0.11
EL-LN	7	4.56	EL-N	7	7.89	0.13
EL-P	6	4.85	EL-P	6	7.11	0.11
EL-LP	5	6.42	EL-LP	5	6.98	0.11
EL-LNP	4	9.10	EL-LNP	4	6.5	0.10
EL-NP	3	9.28	EL-NP	3	6.2	0.10
Norm-mono	2	11.45	Norm-mono	2	1.76	0.08
Normal	1	14.47	Normal	1	0.09	0.02

conditions, the normal sentence was always judged to sound more like NNS than all of the other speech tokens. In such cases, the data collected for the normal sentences can be thrown out and the Kaiser–Serlin method can be applied to the remaining submatrix, but no information can be obtained on the scale value of the normal sentence (it is effectively infinity). Therefore, this study made use of the solution to this problem provided by Krus and Krus (1977), who suggested the following transformation from the proportions, p_{jk}, to the z scores, z_{jk}:

$$z_{jk} = \frac{p_{jk} - p_{kj}}{\sqrt{\frac{p_{jk} + p_{kj}}{N}}}, \qquad (6)$$

where N is the total number of times the stimulus pair, (j,k) was presented. This transformation provides a rational (i.e., finite) z score, even when p_{jk} equals one or zero, which is proportional to the square root of the number of observations. The diagonal entries of the Z matrix are set to zero, the z score of a proportion of .5 (i.e., what would be expected if pairs of the same stimuli were presented). The Kaiser–Serlin method was applied to these transformed scores to produce the scale values. The scale values were then shifted by the amount necessary to set the scale value of the lowest ranked token to zero.

Analysis of visual analog scale data. The distance in centimeters from the left end of the VAS labeled *Not At All Different* was used as an estimate of how different a listener judged a speech token to be from NNS. The lower the number, the less different from normal speech a sentence was judged to be. These distances were used to compute a mean distance for each speech type. A three-way analysis of variance (ANOVA) was performed on the entire VAS-based data set using the three factors: speaker gender, phonemic content, and speech type. Based on the results of this overall analysis, targeted one-way ANOVAs followed by Bonferroni-corrected (Harris, 2001) post hoc t tests were computed (a) on all 10 sentences; (b) on the lowest rated (i.e., least different from normal) EL speech sentence, the normal monotonous speech sentence, and the normal speech sentence; and (c) on the 8 EL speech sentences.

Results

Summary

The average intralistener agreement across all four speaker–sentence conditions (male–voiced, male–voiced/voiceless, female–voiced, female–voiced/voiceless) for the PC task was found to be 88.3% ± 8.9%, using an exact agreement statistic (Kreiman, Gerratt, Kempster, Erman, & Berke, 1993). Intralistener agreement across all four conditions for the VAS task was evaluated using Pearson's r and was found to be .90 ± .09. A summary of the overall results obtained using the PC and VAS procedures across all four speaker-sentence conditions is shown in Table 3. The PC and VAS values for each of the four speaker–sentence conditions are tabulated in Appendixes A and B. The rankings (1–10) of the speech types by the two scaling procedures were almost identical, with the only exception being a simple reversal of the order in which the speech types EL-N and EL-LN appeared on the two scales: EL-N was ranked 8 on the PC scale and 7 on the VAS, whereas EL-LN was ranked 7 on the PC scale and 8 on the VAS. As expected, the two procedures produced scale values having opposite interpretations, with speech types judged to be closer to NNS receiving higher PC scale values and lower VAS values. Also expected were the results that raw EL speech received the worst ranking (most different from NNS) and normal speech received the best ranking (most like NNS). Types of EL speech that combined more than one enhancement were generally ranked as being closer to NNS than EL speech with only one type of enhancement. The sole exception occurred for the pitch-modulated EL speech (EL-P), which had better PC and VAS rankings than EL speech that was enhanced by combining an increase of low-frequency energy with a reduction in the EL-noise (EL-LN).

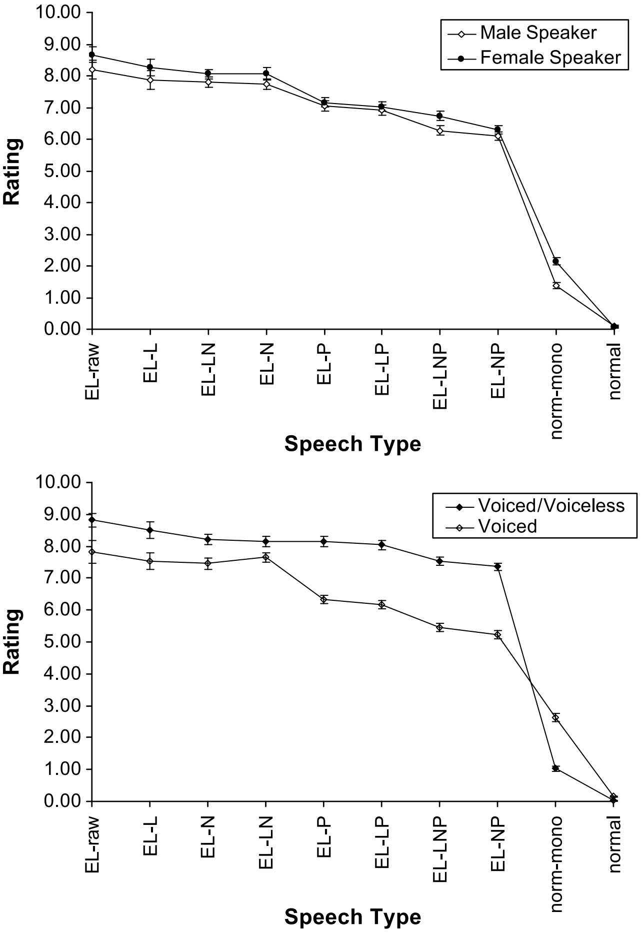
Pitch enhancement had the clearest positive impact on EL speech quality, with pitch modulation being included in all of the four best-ranked types of EL speech. Conversely, increasing the low-frequency content of EL speech seemed to be the least effective enhancement. Not only was it the worst-ranked enhancement when used by itself (only raw EL speech was ranked worse), it was also not included in the best-ranked combination of enhancements (EL-NP was ranked the best). The reduction of the EL-noise, although not as effective as pitch modulation, was clearly more effective than increasing low-frequency content, being included in both of the two best-ranked types of enhanced EL speech. Even though pitch modulation appeared to be the enhancement that had the largest positive impact on EL speech quality, monotonous normal speech, which does not have the proper pitch contour, was still judged to be more like NNS than any version of EL speech.

Statistical Significance

To test whether differences in the ratings of the various types of sentences were significant, we performed a three-way ANOVA on the entire VAS-based data set using the three factors: speaker gender, phonemic content, and speech type. Significant differences were found for all three main effects. There was a significant overall change in scale values across speech types, $F(9, 3462) = 1333$, $p < .0001$, and significantly smaller (closer to NNS) average scale values were found for productions by the male speaker, $F(1, 3470) = 18$, $p < .0001$, and for the all-voiced sentences, $F(1, 3470) = 178$, $p < .0001$.

The interactions between speech type and gender, $F(9, 3462) = 4.1$, $p < .01$, and speech type and phonemic content, $F(9, 3462) = 59.2$, $p < .01$, were also significant. The nature of these interactions is depicted in Figure 2, where the means and standard errors for the VAS ratings of each speech type are plotted separately for the data grouped by gender and phonetic context. Even though the curves for the male and female data (top plot) appear very similar, it is clear the scale values for the male speaker were always slightly lower (better) than the corresponding ratings for the female speaker (except for the normal speech sentence). The differences in average scale values based on phonemic content are much more pronounced (right-hand plot). The all-voiced sentences were always rated better than the sentences comprising both voiced and voiceless sounds, except for the normal-monotonous and normal sentences.

Figure 2. The mean and standard error of the visual analog scale ratings for the different sentence types separated by gender (top) and phonemic content (bottom). Separating the ratings by gender shows that the ratings for the male speaker are consistently lower than those for the female speaker. Separating the data by phonemic content shows that the ratings for the voiced sentence were significantly lower except in the normal monotonous sentence.

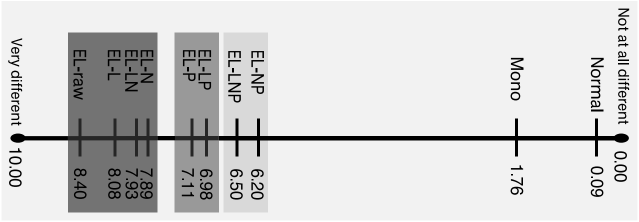


As described in the Method section, an additional one-way ANOVA and post hoc t tests (Bonferroni corrected, $p < .01$) were performed on three different sets of speech types: (a) on all 10 sentences; (b) on the lowest rated (i.e., least different from normal) EL speech sentence, the normal monotonous speech sentence, and the normal speech sentence; and (c) on the 8 EL speech sentences. These tests demonstrated that the ratings for the normal, normal monotonous, and EL-NP speech types were all significantly different ($p < .01$) from each other. Post hoc tests also indicated that (a) the average scale values for the four worst-rated speech types (EL-raw, EL-L, EL-LN, EL-N) did not differ significantly from each other; (b) the average scale values for the EL-P and EL-LP speech types differed significantly from the four worst-rated speech types, but not from each other; and (c) although the EL-NP rating was significantly different from the worst-rated speech types, it was not different from the EL-LNP speech type. Figure 3 illustrates these results by separating the speech types into significantly different groups along the VAS.

Reliability of the Least Squares Estimate

The computed PC-based scale values are only estimates of the true locations of the different speech

Figure 3. This visual representation of the visual analog scale ratings of the different speech types provides a picture of the large gap between the ratings of normal monotonous speech and the best rated version of EL speech (EL-NP). The speech types contained in each gray box were given statistically equivalent ratings. The normal and normal-monotonous speech types were found to be significantly different from each other as well as from all version of EL speech.



stimuli on the psychological continuum of NNS, and thus, it is useful to measure the reliability of these estimates. Kaiser and Serlin (1979) suggested the following measure of reliability:

$$r^2 = \frac{\sum\limits_{i \neq j} \sum (S_i - S_j)}{\sum\limits_{i \neq j} \sum z_{ij}^2}, \quad (7)$$

where S_i, S_j are the computed scale values and the z_{ij} are the measured z-scores discussed in the Method section; r^2 is bounded between 0 and 1, with values closer to 1 indicating a better least squares fit. The r^2 values computed for the different sets of scale values across all speaker–sentence conditions ranged from .854 to .870, with a mean of .863 ± .005, indicating that the least squares model accurately fit the measured data.

The analysis used in this study assumed that Thurstone's Case V (all discriminal dispersions are equal) was applicable to the data. It seemed not unreasonable to make this assumption because the stimuli were all versions of the same sentence. Of the 10 speech types, the one most likely to have a unique discriminal dispersion was normal speech, as it is likely that its presentation would not elicit a great deal of variation on the scale of sounding like NNS. If the discriminal dispersion of normal speech were indeed different, the effect on the overall ranking would be minimal and confined to the scale value of the normal speech (Mosteller, 1951). As the scale values of the normal speech were always found to be considerably greater than those of the other speech types, it is likely that the equal discriminal dispersion assumption had little effect on the scaling results. Furthermore, the high r^2 values suggest that the least squares method, which assumes equal discriminal dispersions, produced an accurate fit to the data.

Discussion

This study sought to better quantify the relative contributions of previously identified acoustic abnormalities to the perception of degraded quality in EL speech by directly evaluating the impact of proposed acoustic enhancements on perceived EL speech quality. The ultimate goal is to use these results to help organize a systematic effort to improve the quality of EL speech for laryngectomy patients.

The results of this study indicate that of the three acoustic abnormalities studied, the lack of pitch information contributes most to the poor quality of EL speech. Next in importance is the EL noise that is generated by the EL. Finally, lack of low-frequency energy appears to contribute least to the degraded quality of EL speech. Thus, in designing an improved EL device, one would gain the most benefit by providing the user with the means to appropriately control pitch while removing, or at least reducing, the amount of EL noise that the EL generates. Although the results of this investigation demonstrate that significant improvement can be made in EL speech quality using the pitch and noise reduction enhancements that were tested, it was also shown that even the best enhanced version of EL speech still has significantly degraded quality when compared to both NNS and monotonous normal speech.

Initially, the PC and VAS results may appear somewhat contradictory with respect to apparent differences in the proximity of EL-NP speech to the normal monotonous speech on the two scales (i.e., EL-NP appears much closer to normal monotonous speech on the PC scale as compared with the VAS). However, much of this difference can be attributed to the inherent operations of the algorithms used to compute these two types of scales. The PC-based scale value of a given stimulus represents where that stimulus falls on the psychological continuum in relation to all of the other stimuli. In fact, if a Z matrix is full, then the scale values are best determined by averaging the columns. Consider the situation of two stimuli, j and k, where both are consistently judged to be "better" than all of the other stimuli, and j is consistently judged to be "better" than k. Although the scale value associated with stimulus j will be greater than that associated with stimulus k, the distance between the scale values is limited by the interaction of the two stimuli in question with all of the other stimuli used in the experiment. Moreover, according to Equation 6, the z scores are proportional to $\sqrt{N}$, where N is the number of observations made. Therefore, a larger number of observations produces larger z scores and, hence, larger scale values. Consequently, because the extent of the scale is limited by the total number of observations that were made and because a relatively large number of stimuli (10) were used in the experiment, the PC-based scale values of EL-NP speech and normal monotonous speech are not as different as they would have been had there been fewer stimuli and/or more observations per stimulus.

On the other hand, the average VAS-based rating for a given speech type is mathematically independent of the ratings for the other speech tokens, and is also not influenced by the number of times each stimulus is rated. Thus, the VAS gave the listeners enough flexibility to, in effect, create an unbalanced scale in which the ordinal ranking of the speech types was the same as it was for the PC-based data, but with a more compressed set of ratings for the EL speech types. However, this interpretation of the results must be tempered by the fact that the VAS can be susceptible to experimental and listener biases, especially the type of logarithmic bias described by Poulton (1989).

These inherent differences between the two perceptual scales underlie why they are both useful and complementary perceptual measures. Because the method of paired comparisons compares each stimulus directly with every other stimulus, it provides a direct means of locating each stimulus on a psychological continuum (in this case, that of sounding most like NNS) based on its relative position on this continuum. Thus, for the purpose of this study, the PC-based scale indicates the relative effectiveness of different combinations of enhancements or, alternatively, the relative importance of different EL speech deficiencies. However, using the Krus and Krus (1977) transformation limits the range of the scale values (since they are proportional to $\sqrt{N}$), thus limiting the ability to fully capture the perceptual distance between the EL speech tokens and both versions of laryngeal speech (normal and static pitch). On the other hand, the VAS is better suited to measure the quality difference between normal laryngeal speech and the different versions of EL speech, because it is based on a direct comparison between each of the speech tokens and the normal speech token. As such, it only provides an indirect measure of the relative speech quality of the speech tokens and, at most, provides a verification of the PC results. Moreover, because only the speech token chosen from the PC task was used for the VAS, the tokens were not all rated an equal number of times. As a result, the differences between the scale values of the highest rated (i.e., the lowest quality), and therefore least frequently judged speech tokens, were not found to be statistically different.

Both the PC and VAS data suggest that the enhancements were not as effective for speech that contains unvoiced phonemes, further limiting the improvement in quality. At least one reason for the overall poorer ratings of the voiced/voiceless EL sentences may be related to the methods that were used to modulate the pitch of these sentences. The pitch contour extracted from the normal speech versions of the voiced–voiceless sentence contained gaps corresponding to the unvoiced parts of the sentence. Before using these extracted contours to enhance EL speech, the pitch values within these gaps were set equal to the last measured pitch value, creating a flat pitch contour for a short period of time. Thus, the pitch contour estimate used in the voiced–voiceless sentence was not as accurate as the one used in the all-voiced sentence, perhaps limiting the effectiveness of the pitch enhancement. However, this reasoning cannot satisfactorily explain the difference in ratings between the non-pitch-enhanced EL speech sentences. All of the EL versions of the voiced–voiceless sentence lacked the proper perceptual cues for unvoiced consonants, a problem inherent in EL speech (Weiss & Basili, 1985; Weiss et al., 1979). It is likely that this missing information contributed to the reduced ratings for the EL versions of the voiced–voiceless sentence.

Although adding pitch information may be the most effective means of improving EL speech quality, and would be required in order to achieve a full reconstruction of an EL user's original voice, it is perhaps the most difficult enhancement to implement because it requires finding a way of estimating what pitch the speaker intends to use, that is, access to underlying linguistic and/or neural processes. The potential difficulties in attempting to use alternative strategies to gain control of pitch are illustrated in a study by Uemi et al. (1994), in which only 2 of the 16 participants studied were able to master the control of an experimental device and thereby produce pitch contours that resembled those in normal speech. In other work by our research group, we have demonstrated the potential feasibility of accessing laryngeal neuromotor signals postlaryngectomy to use in controlling the on/off and pitch of an EL, but this general approach requires further testing and development (Heaton et al., 2004). However, the fact that normal monotonous speech more closely approximated the quality of NNS than any type of EL speech enhancement (including ones with the proper pitch information) provides some hope that EL speech can potentially be significantly improved without having to add pitch information. It also indicates that there are other, unexplored, properties of EL speech that contribute to its unnatural quality. For example, the more limited effectiveness of the three enhancements on EL speech containing unvoiced phonemes suggests that lack of the voiced–voiceless distinction is another important characteristic of EL speech that reduces its quality. This result suggests that a reasonable intermediate goal would be to identify and correct other aberrant acoustic properties of EL speech that enable a closer approximation to normal monotonous speech.

Conclusion

This study sought to better quantify the relative contributions of previously identified acoustic abnormalities to the perception of degraded quality in EL speech: lack of pitch variation, EL noise, and missing low-frequency energy. The results indicate that although EL speech can be most improved by providing proper pitch information and removing the EL noise, the resulting quality is still well below that of NNS or even that of monotonous natural speech. This suggests that, in addition to the widely acknowledged acoustic abnormalities examined in this investigation, there are other attributes that contribute significantly to the unnatural quality of EL speech. Identifying and determining ways to correct for such additional factors is necessary before EL speech can be made to more closely approximate the sound quality of NNS. The results of this investigation form a useful basis for future attempts at improving the quality of EL speech.

Acknowledgments

This study is based on a thesis completed in partial fulfillment of the requirements for the degree of Doctor of Philosophy by the first author in the Speech and Hearing Biosciences and Technology Program at the Harvard-MIT Division of Health Sciences and Technology. This work was supported by grants from the W.M. Keck Foundation and the Department of Veterans Affairs Division of Rehabilitation Research and Development (Grant No. C3343-2DC). We would like to thank Kenneth Stevens of the Research Laboratory of Electronics at MIT, Tom Quatieri of MIT Lincoln Laboratory, and James Kobler of the Voice and Speech Laboratory for their extensive help and expertise.

References

Barney, H. L., Haworth, F. E., & Dunn, H. K. (1959). An experimental transistorized artificial larynx. In B. Weinberg (Ed.), *Readings in speech following total laryngectomy* (pp. 1337–1356). Baltimore: University Park Press.

Cole, D., Sridharan, S., Moody, M., & Geva, S. (1997). Application of noise reduction techniques for alaryngeal speech enhancement. *Proceedings of IEEE TENCON '97. IEEE Region 10 Annual Conference. Speech and Image Technologies for Computing and Telecommunications, 2,* 491–494.

Crystal, T. H., & House, A. S. (1982). Segmental durations in connected speech signals: Preliminary results. *Journal of the Acoustical Society of America, 72,* 705–717.

Diedrich, W., & Youngstrom, K. (1977). *Alaryngeal speech*. Springfield, IL: Charles C Thomas.

Edwards, A. L. (1957). *Techniques of attitude scale construction*. New York: Appleton-Century-Crofts.

Espy-Wilson, C. Y., Chari, V. R., MacAuslan, J. M., Huang, C. B., & Walsh, M. J. (1998). Enhancement of electrolaryngeal speech by adaptive filtering. *Journal of Speech, Language, and Hearing Research, 41,* 1253–1264.

Gates, G. A., Ryan, W., Cantu, E., & Hearne, E. (1982). Current status of laryngectomee rehabilitation: II. Causes of failure. *American Journal of Otolaryngology, 3,* 8–14.

Gates, G. A., Ryan, W., Cooper, J. C., Jr., Lawlis, G. F., Cantu, E., & Hayashi, T., et al. (1982). Current status of laryngectomee rehabilitation: I. Results of therapy. *American Journal of Otolaryngology, 3,* 1–7.

Gray, S., & Konrad, H. R. (1976). Laryngectomy: Postsurgical rehabilitation of communication. *Archives of Physical Medicine and Rehabilitation, 57,* 140–142.

Harris, R. J. (2001). *A primer of multivariate statistics*. Mahwah, NJ: Erlbaum.

Heaton, J., Goldstein, E., Kobler, J., Zeitels, S., Randolph, G., & Walsh, M., et al. (2004). Surface electromyographic activity in total laryngectomees following laryngeal nerve transfer to neck strap muscles: Correlation with vocal and non-vocal behaviors. *Annals of Otology, Rhinology and Laryngology, 109,* 972–980.

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

House, A. S., & Stevens, K. N. (1958). Estimation of formant bandwidths from measurements of transient response of the vocal tract. *Journal of Speech and Hearing Research, 1,* 309–315.

Kaiser, H. F., & Serlin, R. H. (1978). Contributions to the method of paired comparisons. *Applied Psychological Measurement, 2,* 421–430.

King, P. S., Fowlks, E. W., & Peirson, G. A. (1968). Rehabilitation and adaptation of laryngectomy patients. *American Journal of Physical Medicine and Rehabilitation, 47,* 192–203.

Kommers, M. S., & Sullivan, M. D. (1979). Wives' evaluation of problems related to laryngectomy. *Journal of Communication Disorders, 12,* 411–430.

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

Krus, D. J., & Krus, P. (1977). Normal scaling of dominance matrices: The domain-referenced model. *Educational and Psychological Measurement*, *37,* 189–193.

Ma, K., Demirel, P., Espy-Wilson, C., & MacAuslan, J. (1999). Improvement of electrolarynx speech by introducing normal excitation information. *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH), Budapest 1999,* 323–326.

McCree, A. V., & Barnwell, T. P. (1995). A mixed excitation LPC vocoders model for low bit rate speech coding. *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing '95, 3,* 242–250.

Meltzner, G. S. (2003). Perceptual and acoustic impacts of aberrant properties of electrolaryngeal speech. (Doctoral dissertation, Massachusetts Institute of Technology, 2003). *Dissertation Abstracts International, 64,* 4486.

Meltzner, G. S., Hillman, R. E., Heaton, J., Houston, K., Kobler, J. B., & Qi, Y. (2005). Electrolarynx speech: The state-of-the-art and future directions for development. In P. C. Doyle & R. L. Keith (Eds.), *Contemporary considerations in the treatment and rehabilitation of head and neck cancer: Voice, speech, and swallowing.* Austin, TX: Pro-Ed.

Meltzner, G. S., Kobler, J. B., & Hillman, R. E. (2003). Measuring the neck frequency response function of laryngectomy patients: Implications for the design of electrolarynx devices. *Journal of the Acoustical Society of America, 114,* 1035–1047.

Mitchell, H. L., Hoit, J. D., & Watson, P. J. (1996). Cognitive–linguistic demands and speech breathing. *Journal of Speech and Hearing Research, 39,* 93–104.

Morris, H. L., Smith, A. E., Van Demark, D. R., & Maves, M. D. (1992). Communication status following laryngectomy: The Iowa experience 1984-1987. *Annals of Otology, Rhinology and Laryngology, 101,* 503–510.

Mosteller, F. (1951). Remarks on the method of paired comparisons II: The effect of an aberrant standard deviation when equal standard deviations and equal correlations are assumed. *Psychometrika, 16,* 207–218.

Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication, 9,* 453–467.

Poulton, E. C. (1989). *Bias in quantifying judgments*. Hove, U.K.: Erlbaum.

Qi, Y. Y., & Weinberg, B. (1991). Low-frequency energy deficit in electrolaryngeal speech. *Journal of Speech and Hearing Research, 34,* 1250–1256.

Richardson, J., & Bourque, L. (1985). Communication after laryngectomy. *Journal of Psychosocial Oncology, 3,* 83–97.

Sisty, N. L., & Weinberg, B. (1972). Formant frequency characteristics of esophageal speech. *Journal of Speech and Hearing Research, 15,* 439–448.

Stevens, K. N. (1998). *Acoustic phonetics*. Cambridge, MA: MIT Press.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychology Review, 34,* 273–286.

Torgerson, W. S. (1957). *Theory and methods of scaling*. New York: Wiley.

Uemi, N., Ifukube, T., Takahashi, M., & Matsushima, J. (1994). Design of a new electrolarynx having a pitch control function. *IEEE Workshop on Robot and Human,* 198–202.

U.S. Department of Defense. (1999). *Analog-to-digital conversion of voice by 2,400 bit/second mixed excitation linear prediction (MELP)* (MIL-STD-3005). Philadelphia: Author.

Webster, P. M., & Duguay, M. J. (1990). Surgeons' reported attitudes and practices regarding alaryngeal speech. *Annals Otology, Rhinology, and Laryngology, 99,* 197–200.

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

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

Received May 27, 2004

Accepted January 14, 2005

DOI: 10.1044/1092-4388(2005/053)

Contact author: Geoffrey S. Meltzner, now at BAE Systems, 6 New England Executive Park, Burlington, MA 01803. E-mail: geoffrey.meltzner@baesystems.com

Appendix A. Paired comparison scale values based on sentence type and speaker gender.

Male Speaker: Voiced–Voiceless			Male Speaker: All Voiced		
Speech Type	Rank	Scale Value	Speech Type	Rank	Scale Value
EL-raw	10	0	EL-raw	10	0
EL-L	9	0.13	EL-L	9	0.45
EL-P	8	1.21	EL-N	8	1.61
EL-LP	7	1.83	EL-LN	7	1.97
EL-N	6	2.15	EL-P	6	3.13
EL-LN	5	2.5	EL-LP	5	3.98
EL-NP	4	3.76	EL-LNP	4	4.96
EL-LNP	3	4.16	EL-NP	3	5.37
Norm-mono	2	6.04	Norm-mono	2	5.55
Normal	1	6.84	Normal	1	7.42
Female Speaker: Voiced–Voiceless			**Female Speaker: All Voiced**		
EL-raw	10	0	EL-raw	10	0
EL-L	9	0.8	EL-L	9	0.36
EL-N	8	1.97	EL-N	8	1.52
EL-P	7	2.46	EL-LN	7	2.19
EL-LN	6	2.46	EL-P	6	2.91
EL-LP	5	3.22	EL-LP	5	3.8
EL-NP	4	4.29	EL-LNP	4	4.43
EL-LNP	3	4.65	Norm-mono	3	5.01
Norm-mono	2	6.31	EL-NP	2	5.14
Normal	1	7.38	Normal	1	7.29

Appendix B. Visual analog scale values separated by sentence type and speaker gender.

Male Speaker: Voiced–Voiceless Sentence					Male Speaker: All Voiced Sentence				
Speech Type	Rank	Rating	s_{m}	*N*	Speech Type	Rank	Rating	s_{m}	*N*
EL-raw	10	8.58	0.29	26	EL-LN	10	7.81	0.21	57
EL-LP	9	8.23	0.18	67	EL-L	9	7.70	0.36	23
EL-P	8	8.11	0.25	53	EL-N	8	7.51	0.25	49
EL-L	7	8.00	0.42	29	EL-raw	7	7.40	0.65	13
EL-N	6	7.90	0.21	74	EL-P	6	6.36	0.19	83
EL-LN	5	7.80	0.22	82	EL-LP	5	6.06	0.17	102
EL-LNP	4	7.34	0.18	119	EL-LNP	4	5.26	0.18	124
EL-NP	3	7.26	0.19	110	EL-NP	3	5.14	0.17	133
Norm-mono	2	0.57	0.06	161	Norm-mono	2	2.34	0.15	137
Normal	1	0.04	0.03	179	Normal	1	0.18	0.05	179
Female Speaker: Voiced–Voiceless Sentence					**Female Speaker: All Voiced Sentence**				
EL-raw	10	9.25	0.26	15	EL-raw	10	8.15	0.38	17
EL-L	9	8.95	0.33	33	EL-LN	9	7.50	0.22	66
EL-N	8	8.61	0.26	59	EL-N	8	7.42	0.26	51
EL-LN	7	8.57	0.21	70	EL-L	7	7.36	0.36	25
EL-P	6	8.17	0.18	70	EL-P	6	6.29	0.20	82
EL-LP	5	7.90	0.20	87	EL-LP	5	6.29	0.19	102
EL-LNP	4	7.74	0.17	119	EL-LNP	4	5.69	0.19	116
EL-NP	3	7.45	0.18	111	EL-NP	3	5.32	0.17	132
Norm-mono	2	1.50	0.13	156	Norm-mono	2	2.93	0.22	129
Normal	1	0.01	0.00	180	Normal	1	0.11	0.02	180