Perceived naturalness of electrolaryngeal speech produced using sEMG-controlled vs. manual pitch modulation

K.F. Nagle [1,2], *J.T. Heaton*[2]

[1] Department of Speech-Language Pathology, Seton Hall University, U.S.A.
[2] Center for Laryngeal Surgery & Voice Rehabilitation, Massachusetts General Hospital, U.S.A.
naglekat@shu.edu, james.heaton@mgh.harvard.edu

Abstract

Producing speech with natural prosodic patterns is an ongoing challenge for users of electrolaryngeal (EL) speech. This study describes speech produced using a method currently in development, wherein a prosodic pattern is derived from skin surface electromyographical (sEMG) signals recorded from under the chin (submental surface).

Eight laryngectomees who currently use a TruTone™ EL as their primary or backup mode of speech provided samples of EL speech in two modes: conventional thumb-pressure pitch-modulated control (represented by the TruTone™ EL; Griffin Laboratories, CA, U.S.A.) and sEMG-based pitch-modulated control (EMG-EL). Ratings of perceived naturalness were obtained from ten listeners unfamiliar with EL speech.

Listener ratings indicated that five speakers produced equally natural speech using both devices, and three produced significantly more natural speech using the EMG-EL than the TruTone™ EL. Mean fundamental frequency (f0) was similar within speakers for both modes; however, mean f0 range and standard deviation were significantly larger for the EMG-EL than for the TruTone™ EL, despite both devices having similar potential f0 range. This study showed that the EMG-EL provides an intuitive means of controlling f0-based prosodic patterns that are more natural-sounding than push-button control for some EL users.

Index Terms: alaryngeal speech, electrolaryngeal speech, fundamental frequency, laryngectomy, naturalness, prosody

1. Introduction

Traditional electrolaryngeal (EL) speech is monotonic and robotic-sounding, lacking natural prosodic control. A few currently available devices, such as the TruTone™ EL (Griffin Labs, CA, U.S.A.), provide pitch modulation via haptic pressure; however, few users actually utilize this capability [1, 2]. Prior studies have demonstrated that an electromyographic (EMG) interface can be effective for controlling dynamic fundamental frequency (f0) variation for EL speech (EMG-EL) [3, 4, 5, 6]. In this study we evaluated perceived speech naturalness of speech produced using the newest version of the EMG-EL versus a TruTone™ EL.

1.1. EMG-Controlled Electrolarynx

The EMG-EL system allows EMG-based or manual control of a customized handheld EL, with multiple possible control combinations. A wireless, battery-powered sensor filters and amplifies EMG signals from the neck surface, generates an EMG envelope, and transmits it to a modified TruTone™ EL for voice onset, offset and proportional f0 (pitch) modulation [7]. In this way, speech with dynamic f0 can be produced with hands-free control, although the EL itself must still make contact with a 'sweet spot' for sound transmission through neck, chin or face surface [8].

EMG-based EL control was investigated by Stepp and colleagues [6] for speech produced by eight laryngectomees using an earlier version of the current EMG-EL hardware and software. Of the seven neck and face sEMG electrode placement locations studied, the best EMG-EL control was obtained from the superior ventral neck and submental surfaces. Listener assessment of structured sentences and spontaneous speech when using 1) the EMG-EL, 2) a conventional EL, and 3) normal laryngeal speech indicated that both EL devices were significantly less natural-sounding than laryngeal speech, and that the two EL devices did not differ from one another in supporting speech naturalness. The lack of an advantage for EMG-based f0 modulation was unexpected given the known importance of dynamic f0 in listener assessment of speech naturalness [9]. The authors suggested that the proportional relationship between the EMG envelope and resulting f0 should have been set in the EMG-EL system to provide a greater f0 range than had been used. Specifically, the normal (laryngeal) naturalness exemplar used in that study had a speaking f0 range of 79.4 Hz and SD f0 of 16.9 Hz, whereas mean f0 range for sentences produced with the EMG-EL device was 20.8 Hz, with mean SD f0 4.1 Hz.

Inexperienced listeners in that study also identified "melodic/lots of intonation" as a predominant quality demonstrated by the most natural voices, further suggesting that a greater EMG-EL f0 range would have yielded greater naturalness ratings. The newest version of the EMG-EL system provides a real-time visual display of the EMG envelope used for proportional f0 control on a computer monitor, and allows the operator to set the f0 range in relation to the envelope magnitude. For example, by setting the EL f0 minimum and maximum in relation to the EMG envelope at baseline (rest) versus strong speaking effort (respectively), we hope to provide a more dynamic and naturalistic f0 range than has been used in prior studies with earlier versions of the EMG-EL system.

Based on findings by Stepp et al [6], we placed the wireless sEMG sensor on the submental surface (under the chin) on the opposite side of the point of EL contact for each Speaker in this study (Figure 1). Although the EL of the EMG-EL system can be mounted on the neck for hands-free operation, in this study we chose to have Speaker participants hold the EL and operate

its onset and offset as they would their own EL (i.e. button activation), so that speaking conditions would be as similar as possible for hand-held use of the EMG-EL versus TruTone™ voice prostheses. Therefore, the main difference under study between these devices was in f0 modulation control: sEMG energy from submental muscles for the EMG-EL vs. pressure-sensitive manual control for the TruTone™ EL. Acoustic and perceptual characteristics of speech produced using manual (TruTone™) and submental sEMG (EMG-EL) pitch modulation are compared in this study for experienced TruTone™ EL users.

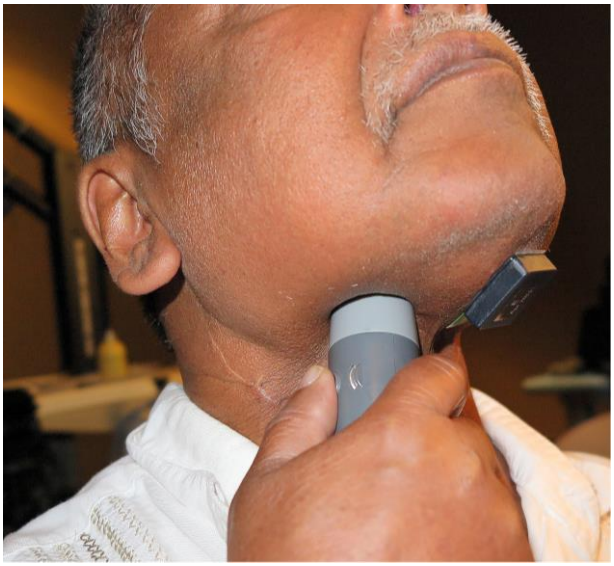

Figure 1. *Positioning of wireless sEMG sensor relative to Speaker's EL sweet spot for sound transmission.*

1.2. Hypotheses

1.2.1. Speech produced using EMG-EL pitch modulation will be rated by inexperienced Listeners as more natural than speech produced using thumb-button pitch modulation.

1.2.2. The range and standard deviation of fundamental frequency control will be greater for EMG-EL pitch modulated speech than for thumb-button pitch modulation.

2. Methods

Eight male alaryngeal Speakers who currently use a TruTone™ EL as either their primary (n=6) or backup mode of communication (n=2), provided recordings of speech using the TruTone™ EL and the EMG-EL. A hardware description of the EMG-EL system is found in [7]. Pitch range (f0max – f0min) and variation (SD f0) within each utterance were obtained. Ten Listeners unfamiliar with alaryngeal speech provided ratings of perceived naturalness.

2.1. Stimuli

EMG signals emanating from the submental surface inherently correspond to vocal effort and intonation, providing intuitive pitch control of the EMG-EL. For this reason, Speakers received only brief operational instructions before recording. They produced two recordings of each of the following sentences, using each device (2 x 4 Sentences x 2 Modes = 16 stimuli per Speaker):

1. "His sister Mary and brother George went along, too."
2. "She filled the bag with tomatoes."
3. "Try to work things out."
4. "You can see that they didn't have far to go."

Stimuli were recorded in a quiet room at a sampling frequency of 44.1 kHz with 16-bit quantization, using a portable audio recorder (Tascam DR-40 Linear PCM recorder) and headset microphone (AKG C520 MicroMic) with a mouth-to-microphone distance of 5 cm.

The f0 range of the TruTone™ EL was 62-155 Hz based on push-button pressure from activation to maximal depression. The f0 range of the EMG-EL was 50-160 Hz, proportional to the sEMG envelope. The relationship between f0 and sEMG was set individually for each Speaker using a computer interface. The EMG-EL f0 would begin to increase from minimum when the sEMG envelope exceeded 8% of the resting baseline, and would achieve maximum when the sEMG envelope reached a maximal voluntary contraction level (identified during tongue protrusion). Therefore, the relationship between sEMG envelope and f0 across the potential f0 range differed for each speaker.

2.1.1. Stimulus Preparation

Stimuli were saved as .wav files and amplitude normalized to 70 dB SPL using a customized Praat script [10]. Each stimulus was prepared with 50 ms of silence preceding the onset and following the offset of speech.

2.2. Measurement

2.2.1. Perceptual Data

Ten native English-speaking Listeners (2 male, age M=25.3, SD =3.02) passed a hearing screening at 25dB for the octave frequencies between 250 and 4,000 Hz.

Stimuli were presented as individual icons on a laptop computer screen using a customized software program (Matlab; MathWorks, Natick, MA) designed for visual sorting and rating of stimuli [11, 12]. Listeners wore headphones (Sony MDR 506) and were instructed to adjust the loudness of stimulus presentation to a comfortable level. Stimuli could be played and replayed by clicking on an icon (Figure 2).

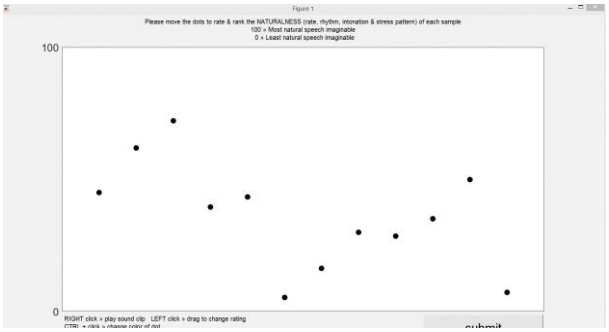

Figure 2: *Example screenshot of the graphic user interface for rating perceived naturalness.*

Listeners were given the following instructions before completing a short practice round of rating:

"You will be rating the NATURALNESS of samples of electrolaryngeal speech. None of these samples will sound completely natural; however, some will probably sound better than others. Speech samples have been defined as "natural" if they conform to the listener's standards of rate, rhythm, intonation and stress pattern. Using this program you will be able to RANK the sample in order of naturalness and RATE them on a scale of 0-100."

As shown in Figure 2, Listeners sorted the stimuli into those that were relatively more or less natural by dragging the icons to the upper or lower part of the computer screen. They then rated the samples by placing each icon on a 100 mm vertical visual analog scale (VAS). In this way they indicated not only rank but approximate degree of difference between each sample in terms of perceived naturalness.

Stimuli were presented in 12 sets of 12. Each set contained 10-11 unique stimuli plus 1-2 stimuli repeated from another set (so that measures of intra-rater agreement could be obtained). Listeners heard all 12 sets (N = 128 unique samples + 16 repeated samples), with breaks between sets as needed. Completion of the study took 60-90 minutes.

2.2.2. Acoustic Data

Acoustic analysis was conducted using Praat (Boersma & Weenink, 2015), with f0 sampling at 2 kS/s. For comparison with perceptual measures, only summary measures of intonation were taken. The mean and standard deviation of f0 across each sample was calculated based on zero crossings between the onset and offset of EL "voicing" on waveform and spectrogram displays. Likewise, f0 range for each speech sample was calculated by manual identification of the zero crossings between the onset and offset of EL "voicing" on waveform and spectrogram displays.

2.3. Data Analysis

For Hypothesis 1.2.1, mean ratings of perceived naturalness were computed for each stimulus and analyzed using 3-factor ANOVA (Speaker x Sentence x Mode), including the interaction between Speaker and Mode. Post hoc t-tests comparing naturalness ratings between Modes for each Speaker were run with Bonferroni corrections ($\alpha = .00625$).

For Hypothesis 1.2.2, 2-factor ANOVA (f0 range x SD f0) was used to measure pitch variability for EMG-EL compared to TruTone™ speech ($\alpha = .05$).

2.3.1. Rater Reliability & Agreement

Rater reliability was established using intraclass correlation coefficients (ICCs). Inter-rater reliability was calculated using a 2-way mixed model in which raters were considered a fixed factor [13]. The average measures ICC (2, *k)* provides the mean reliability for all ratings based on the relation between a single rating and the mean of all ratings for that sample. Raters were highly consistent, with ICC (2,*k*)=.899 (95% CI [.862, 929]).

Intra-rater agreement was calculated for each Listener using a one-way random effects model (1, *k*) for the first and second ratings of individual Listeners for repeated stimuli (n=16). Repeated ratings agreed strongly, with mean ICC (1,*k*) = .92 (range=.82-.99).

3. Results

3.1. Perceived Naturalness

All Speakers' productions were rated as natural or more natural for EMG-EL speech than for TruTone™ speech. Table 1 shows descriptives and results of 3-way ANOVA for naturalness ratings, and Figure 3 displays within-Speaker differences in naturalness ratings between speech Modes. Notably, Speaker 4, who had the highest naturalness ratings, is the most expert TruTone™ user ever encountered by the authors. Consistent with this observation, two Listeners asked if he was a laryngeal speaker during the rating task.

Table 1. *Descriptives and results (3-way ANOVA) for naturalness (mm), by Speaker and Mode.*

	Mode				
	EMG-EL			TruTone™	
Condition	*M*	*(SD)*		*M*	*(SD)*
Speaker					
1	20.64	(3.94)		9.83	(3.95)
2	37.85	(6.69)		33.08	(10.63)
3	42.55	(8.46)		32.00	(7.73)
4	60.61	(5.97)		65.11	(12.14)
5	39.46	(9.68)		23.35	(5.41)
6	46.60	(7.35)		36.75	(6.17)
7	38.77	(15.27)		36.76	(8.60)
8	31.05	(4.86)		26.12	(5.84)
Model Results	*F*	*(df)*	*p*	ω^2	
Main Effects					
Speaker	42.69	(7,109)	.000	0.64	
Mode	21.95	(1,109)	.000	0.04	
Sentence	0.97	(3,109)	.408	0.00	
Interactions					
Mode*Speaker	2.40	(7,109)	.025	0.02	

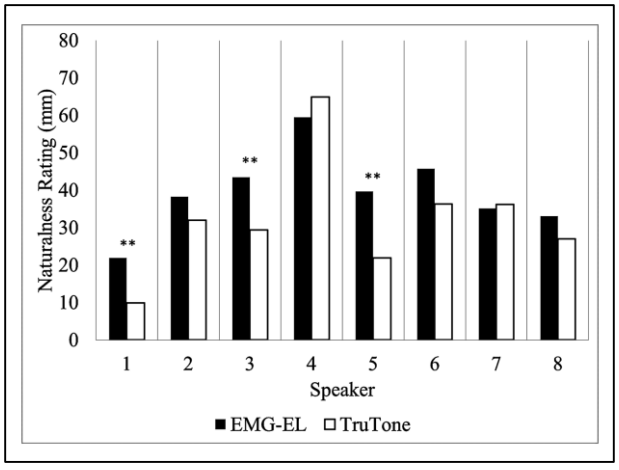

Figure 3. *Mean naturalness ratings for each Speaker by Mode. (** p < .00625 with Bonferroni correction.)*

For Speakers 2, 4, 6, 7 and 8, both speech Modes achieved similar naturalness ratings ($p>.00625$). EMG-EL speech produced by Speakers 1, 3 and 5 was rated significantly more natural than TruTone™ speech ($p<.00625$).

3.2. Acoustic Characteristics

3.2.1. f0 Range

Mean f0 range was significantly greater for speech produced in EMG-EL Mode than TruTone™ EL Mode, for all Speakers ($p < .001$), as shown in Table 2.

Table 2. *Mean f0 range in Hz, by Speaker.*

Speaker	EMG-EL	TruTone™
1	53.40	22.56
2	43.82	11.92
3	47.35	8.37
4	57.52	26.37
5	53.30	6.47
6	59.65	21.17
7	42.39	9.74
8	51.06	15.23
Mean	51.15	15.32

3.2.2. Mean & SD f0

Speech was produced at roughly the same mean f0 across Modes for each Speaker (Table 3). Pitch variation, as measured by standard deviation of f0, was significantly higher when using the EMG-EL than TruTone™ EL, for all Speakers ($p<.001$).

Table 3. *Mean f0 in Hz per Speaker, by EL Mode.*

Speaker	EMG-EL *M(SD)*	TruTone™ *M(SD)*
1	119 (13)	135 (6)
2	90 (10)	81 (2)
3	71 (12)	79 (1)
4	93 (15)	80 (6)
5	78 (15)	71 (1)
6	95 (17)	89 (4)
7	68 (13)	70 (2)
8	97 (15)	90 (8)
Mean	89 (14)	90 (3)

Speaker 1, whose naturalness ratings were consistently low, expressed a preference for higher pitch and produced a notably higher f0 from other Speakers throughout the testing.

4. Discussion

Results of this study indicate that the EMG-EL is a viable method of providing natural-sounding pitch modulation for users of EL devices. All Speakers owned and had at least some experience with the TruTone™ device, but none, including the exceptional TruTone™ EL user (Speaker 4), achieved significantly greater naturalness with the TruTone™ EL than the EMG-EL. Conversely, all Speakers had produced only a few minutes of speech with the EMG-EL, yet all achieved roughly equal or better naturalness ratings with the EMG-EL than with the TruTone™ EL.

As predicted, simple acoustic measures of EL speech in this study indicated greater pitch variability produced by these new EMG-EL users than for even the most experienced TruTone™ Speaker participants. This was true even when accounting for the 17 Hz greater potential f0 range for the EMG-EL compared to the TruTone™ EL.

Given that all Speakers produced significantly greater pitch variability for EMG-EL speech, but that only three had significantly greater naturalness ratings, future research should better characterize the relationship between f0 range and speech naturalness within and across individuals. In the present study we set the f0 minimum and maximum intuitively at 8% above baseline and at sustained maximum voluntary contraction (respectively), but it is unknown whether this provides an optimal relationship between EMG and f0. Moreover, the EMG-EL system has several available EMG envelope low-pass filter settings (i.e. determining the rate of envelope change). Although we chose an envelope speed that supports the most natural-sounding speech to our ears, future work should systematically study the perceptual impact of different linear and nonlinear rates of EMG-based f0 change during speech.

It is also likely that factors other than pitch variability contribute to naturalness, even for EL speech. Rate, rhythm, stress pattern and volume are just a few of these known factors [14,15]. Although amplitude was equalized for all samples in this study, we did not control speech rate, rhythm or stress patterns when eliciting stimuli. Speakers were merely encouraged to be natural. Speaker 4, who had the highest naturalness ratings, was observed to pace his speech differently from the others. For example, he inserted a measurable pause into one of his recordings, which may have improved the naturalness of his speech rhythm. Speaker 1, on the other hand, had much higher pitch than the rest of the Speakers in this study, and his naturalness ratings were consistently low. Given the effect of mean pitch on ratings of acceptability and gender identification in previous studies of alaryngeal speech [16], it is critical to find the appropriate mean f0 for individual Speakers.

5. Conclusions

Pitch modulation is an important aspect of speech naturalness lost when laryngectomees speak using any of several monotonic EL models. When pitch modulation is available through push-button control (e.g. the TruTone™ EL), it is often underutilized through lack of skill and/or inappropriate hardware settings. Naturalistic f0 modulations can potentially be provided through EMG-based f0 control (EMG-EL), where f0 changes proportionally to an EMG envelope obtained from the neck surface.

In this report we show that speech produced using the EMG-EL voice prosthesis was equally or more natural sounding compared to speech using the TruTone™ EL. This was true despite the fact that Speaker participants in this study were all TruTone™ EL owners and also had very little experience using the EMG-EL (< 1 hour). Future work is needed to determine an optimal relationship between vocal-related EMG and the resulting f0 range and rate of change during EMG-EL speech.

6. Acknowledgements

The authors thank Cara Stepp, PhD for script used to obtain naturalness ratings, members of the International Association of Laryngectomees for accommodating data collection, and Griffin Laboratories for providing all EL hardware used in this study. This work was supported by R42DC011212-02 to Mark Robertson and James Heaton (multiple PI).

7. References

[1] R. L. Keith, J. C. Shanks, and P. C. Doyle, "Historical highlights: Laryngectomy rehabilitation," in *Contemporary Considerations in the Treatment and Rehabilitation of Head and Neck Cancer: Voice, Speech, and Swallowing*, Austin, TX: Pro-Ed, 2005, pp. 17–58.

[2] Y. Saikachi, "Development, perceptual evaluation, and acoustic analysis of amplitude-based F0 control in Electrolarynx speech," Thesis, Massachusetts Institute of Technology, 2009.

[3] E. A. Goldstein, J. T. Heaton, J. B. Kobler, G. B. Stanley, and R. E. Hillman, "Design and implementation of a hands-free electrolarynx device controlled by neck strap muscle electromyographic activity," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 2, pp. 325–332, 2004.

[4] E. Goldstein, J. Heaton, C. Stepp, and R. Hillman, "Training effects on speech production using a hands-free electromyographically controlled electrolarynx," *Journal of Speech, Language & Hearing Research*, vol. 50, no. 2, pp. 335–351, 2007.

[5] H. L. Kubert, C. E. Stepp, S. M. Zeitels, J. E. Gooey, M. J. Walsh, S. R. Prakash, R. E. Hillman, and J. T. Heaton, "Electromyographic control of a hands-free electrolarynx using neck strap muscles," *Journal of Communication Disorders*, vol. 42, no. 3, pp. 211–225, 2009.

[6] C. E. Stepp, J. T. Heaton, R. G. Rolland, and R. E. Hillman, "Neck and face surface electromyography for prosthetic voice control after total laryngectomy," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 17, no. 2, pp. 146–155, 2009.

[7] J. T. Heaton, M. Robertson, and C. Griffin, "Development of a wireless electromyographically controlled electrolarynx voice prosthesis," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2011, pp. 5352–5355.

[8] P. C. Doyle, "Clinical procedures for training use of the electronic artificial larynx," in *Contemporary Considerations in the Treatment and Rehabilitation of Head and Neck Cancer*, Austin, TX: Pro-Ed, 2005, pp. 545–570.

[9] G. S. Meltzner and R. E. Hillman, "Impact of aberrant acoustic properties on the perception of sound quality in electrolarynx speech," *Journal of Speech, Language & Hearing Research*, vol. 48, no. 4, pp. 766–79, 2005.

[10] P. Boersma and D. Weenink, *Praat: Doing phonetics by computer*. 2015.

[11] S. Granqvist, "The visual sort and rate method for perceptual evaluation in listening tests," *Logoped Phoniatr Vocol*, vol. 28, no. 3, pp. 109–116, 2003.

[12] S. Anand and C. E. Stepp, "Listener perception of monopitch, naturalness, and intelligibility for speakers with Parkinson's disease," *Journal of Speech Language and Hearing Research*, vol. 58, no. 4, p. 1134, Aug. 2015.

[13] K. O. McGraw and S. P. Wong, "Forming inferences about some intraclass correlation coefficients," *Psychological Methods*, vol. 1, no. 1, pp. 30–46, 1996.

[14] M. Klopfenstein, "Relationship between acoustic measures and speech naturalness ratings in Parkinson's disease: A within-speaker approach," *Clinical Linguistics & Phonetics*, Sep. 2015.

[15] K. Yorkston, D. R. Beukelman, E. A. Strand, and K. R. Bell, *Management of motor speech disorders in children and adults*, 2nd ed. Austin Tex.: Pro-Ed, 1999.

[16] K. F. Nagle, T. L. Eadie, D. R. Wright, and Y. A. Sumida, "Effect of fundamental frequency on judgments of electrolaryngeal speech," *American Journal of Speech-Language Pathology*, vol. 21, no. 2, pp. 154–166, 2012.