Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 78 (2016) 34–41

www.elsevier.com/locate/specom

Generating tonal distinctions in Mandarin Chinese using an electrolarynx with preprogrammed tone patterns

Liana Guo[a,b], Kathleen F. Nagle[b], James T. Heaton[a,b,*]

[a] *Massachusetts General Hospital Institute of Health Professions, 36 1st Ave, Boston, MA 02129, United States*

[b] *Massachusetts General Hospital Center for Laryngeal Surgery and Voice Rehabilitation, 1 Bowdoin Sq, 11th Floor, Boston, MA 02114, United States*

Received 1 July 2015; received in revised form 7 January 2016; accepted 13 January 2016
Available online 21 January 2016

Abstract

An electrolarynx (EL) is a valuable rehabilitative option for individuals who have undergone laryngectomy, but current monotone ELs do not support controlled variations in fundamental frequency for producing tonal languages. The present study examined the production and perception of Mandarin Chinese using a customized hand-held EL driven by computer software to generate tonal distinctions (tonal EL). Four native Mandarin speakers were trained to articulate their speech coincidentally with preprogrammed tonal patterns in order to produce mono- and di-syllabic words with a monotone EL and tonal EL. Three native Mandarin speakers later transcribed and rated the speech samples for intelligibility and acceptability. Results indicated that words produced using the tonal EL were significantly more intelligible and acceptable than those produced using the monotone EL.

Keywords: Electrolarynx; Mandarin Chinese; Tonal control; Speech intelligibility.

1. Introduction

Each year, thousands of individuals undergo a total laryngectomy, a standard surgical treatment for advanced laryngeal cancer that results in complete removal of the larynx and leaves them without the ability to phonate normally. Fortunately, several voice prostheses, including the commonly used electrolarynx (EL) provide an alternative source of verbal communication for these individuals. To produce speech, the EL transmits electromechanical vibrations, which can then be shaped by movements of the articulators, through the neck tissue. Due to its portability, ease of use, and readiness to serve as a backup when experiencing difficulty with other modes of alaryngeal speech, the EL is chosen by more than half of Laryngectomees as their primary mode of communication (Hillman et al., 1998).

Although the EL requires little training for users to achieve voicing, most ELs generate little, if any, pitch variation during phonation, contributing to a robotic, unnatural speech quality. Conventional ELs that vibrate on a single fundamental frequency (F_0) have been shown to create a particular deficit in the speech intelligibility of tonal languages such as Thai, Mandarin, and Cantonese (Gandour et al., 1988; Liu et al., 2006; Ng et al., 1998). For words to be correctly perceived in a tonal language when context clues are lacking, a listener needs to hear not only the speech sounds, but also the underlying pitch changes within each syllable (tones). In Mandarin, each syllable contains one of four basic tones (plus a fifth, neutral tone) that make use of F_0 to differentiate the meaning of words with the same sound pattern. Tone 1 has high-level (HL) pitch, Tone 2 middle-rising (MR) pitch, Tone 3 falling-rising (FR) pitch, and Tone 4 high-falling (HF) pitch. For example, the syllable "ma," produced with the four tones, means "mother," "numb," "horse," and "scold," respectively. Liu et al. (2006) found that speakers using a conventional, monotone EL produced F_0 contours that were invariably level. Consequently, listener identification of tones

* Corresponding author at: Massachusetts General Hospital Center for Laryngeal Surgery and Voice Rehabilitation, 1 Bowdoin Sq, 11th Floor, Boston, MA 02114, United States.
E-mail address: james.heaton@mgh.harvard.edu (J.T. Heaton).

http://dx.doi.org/10.1016/j.specom.2016.01.002
0167-6393/

was significantly poorer than those produced by normal laryngeal (NL) speakers.

Studies have found that compared to other phonetic components, including duration and amplitude, F_0 contour provides the most important cue for tone perception for tonal languages (Whalen & Xu, 1992; Zhang, Qi, Song & Liu, 1981). Therefore, ELs that provide dynamic pitch variation can potentially better serve the communicative needs of Mandarin EL speakers. Current EL devices have the capacity to modulate F_0 through several methods: applying varying amounts of finger pressure on a single button (e.g. Western Electric #5, Western Electric; TruTone EL, Griffin Laboratories); controlling expiration pressure from the neck stoma (Uemi et al., 1994); filtering electromyographic (EMG) signals obtained from neck muscle contractions (Goldstein et al., 2004); and adjusting forearm tilt movement (Matsui et al., 2013). These devices have allowed speakers to convey natural intonation patterns in English and Japanese with varying degrees of success; however, effects on the vocal rehabilitation of tonal languages have not yet been shown.

Whereas F_0 contours of intonation can occur over the course of several seconds, tonal contours typically span milliseconds (Xu, 1997). Effective application of these EL devices to Mandarin requires the ability to modulate F_0 rapidly to successfully produce the four tones. To achieve optimal intelligibility, pitch contours generated by the ELs would also need to closely match the typical shape and frequency values for each tone. Considering these factors, the aforementioned EL devices with real-time pitch control have limitations that may fall short of normal F_0 control. Using expiration or finger pressure to provide precise pitch control is difficult to master and may reduce normal speaking rate (Liu and Ng, 2007), and the responsiveness of the EMG-EL's low-pass filter appears to be too slow to support tone production. Additionally, the fixed initial F_0 settings of the TruTone EL and EMG-EL make it difficult to generate appropriate starting tones and F_0 height for each tone.

Wan et al. (2012) presented a viable EL option for Mandarin using the movement of a trackball to control pitch (WT-EL). Users were required to manipulate the trackball with their thumb by moving the direction of the trackball to reflect the differences in tones during phonation. They found that, compared to the monotone EL, the WT-EL performed significantly better in measures of perceptual accuracy and acceptability. However, technical limitations of the device (i.e., 100 ms required to reset to initial F_0) made it difficult for users to consistently generate tonal contours similar to normal speech and produce continuous speech. Additionally, users may find the trackball difficult to master since they needed adequate hand control to precisely manipulate the trackball.

To further enhance Mandarin EL speech, ELs need to closely approximate natural tonal contours without sacrificing convenience. The present study explores the feasibility of achieving EL tonal control for Mandarin using an EL controlled by computer software to generate tonal distinctions. Speakers were trained to articulate their speech coincidentally with preprogrammed pitch patterns to produce words in Mandarin. Use of preset pitch patterns allowed for generated tones to resemble those produced by the NL voice while reducing variability caused by individual speaker differences. This study examined the production and perception of Mandarin using a tone-capable (tonal) EL compared to a monotone EL.

Fig. 1. A TruTone® EL modified to receive an input sine wave determining the instantaneous fundamental frequency (F_0) of the EL voice. Quarter provides size reference.

2. Methods

2.1. Tonal EL design

A TruTone® electrolarynx (EL) was modified by Griffin Laboratories (Temecula, CA) to receive an input signal that determined vocal fundamental frequency (F_0). The TruTone® normally detects the amount of pressure applied to its activation button and modulates F_0 proportionally. The modified TruTone® was physically altered to receive an external electrical signal through a custom input channel (see the black wire in Fig. 1) instead of using the pressure-sensitive activation button. The circuitry within the TruTone® was reprogrammed to detect the instantaneous frequency of a sine wave input and drive its mechanical transducer (e.g. sound source) at the same frequency. Sine wave patterns matching the four Mandarin tones or a monotone condition (see Section 2.2) were synthesized using audio editing software (Adobe Audition CS6, Adobe Systems), and inputted into the TruTone® through the audio output of a personal computer. To produce speech, users depressed the EL activation button to initiate a ready state, began playback of a preset tone on

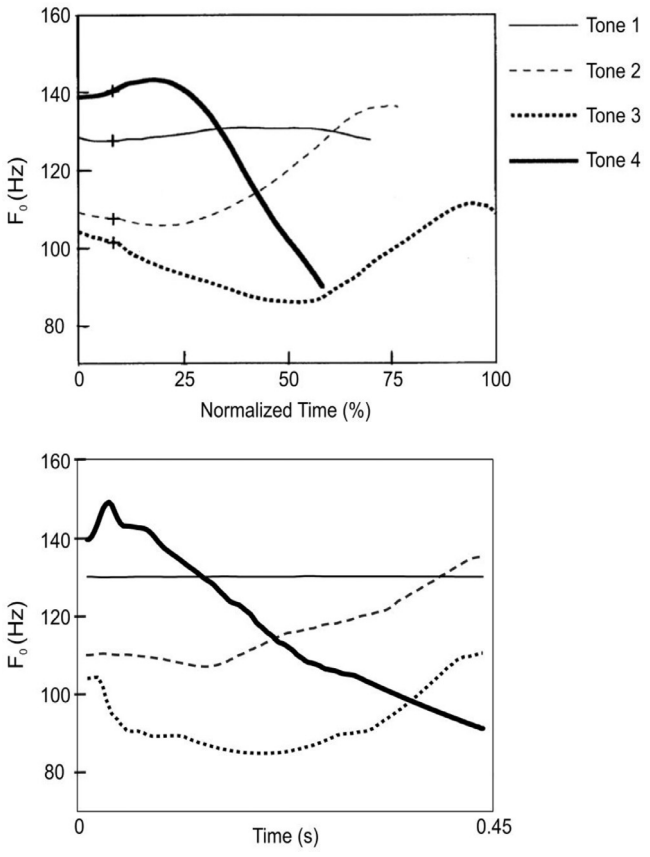

Fig. 2. Comparison of tonal contours in Mandarin for NL speakers and computer software-generated tones. The top panel is a reproduction of Figure 2 from Xu (1997) depicting average tonal contours of eight male speakers when producing the syllable /ma/. The bottom panel shows preprogrammed lexical tones used as the driving signal for the tonal EL fundamental frequency (F_0). The tonal EL produced these F_0 patterns consistently across different speakers and syllables. The monotone F_0 condition at 108 Hz is not shown.

the computer, and moved their articulators to speak as they normally would during EL voicing (while suppressing their natural phonation). The EL could be used in a tonal mode (tonal EL) or monotone mode (monotone EL) depending on the frequency pattern of the input sound file.

2.2. Tone programming

Monosyllabic and disyllabic words were chosen for the study to minimize the effect of contextual cues on intelligibility judgments and to provide simple speech tasks for coordinating the timing of articulation with preprogrammed EL F_0 patterns. Tones for monosyllabic words, based on typical F_0 patterns for each tone (Xu, 1997), were generated using Adobe Audition at 16-bits/sample at 44,100 Hz (see Fig. 2). The monotone sound file was set to 108 Hz, based on the average F_0 of Chinese males documented by Traunmüller and Eriksson (Unpublished results). To control for duration effects on listener identification of tones, duration for all tones was conformed to the typical length of the longest tone, FR. Reports of isolated tone production range from approximately 350 ms to 450 ms (Whalen and Xu, 1992; Xu, 1997); therefore sound files for all tones were created in that range, in intervals of 25 ms (i.e. 350 ms, 375 ms, 400 ms, 425 ms, 450 ms). Two native speakers of Mandarin later listened to the five different syllable durations and concluded that 450 ms seemed to be the most appropriate for producing monosyllabic words with the EL.

Tones for disyllabic words were created by combining the tones to produce sound files that were approximately 900 ms in duration. Several adjustments to tonal contours of the FR tone in disyllabic sequences were also made based on rules for tones produced in succession, as discussed in Xu's (1997) research. Two native Mandarin speakers judged tones produced with the EL to be perceptually accurate and at an acceptable rate to produce disyllabic words.

2.3. Participants

2.3.1. Speakers

Six male native Mandarin Chinese speakers with normal neck anatomy were invited to participate in the study. Recordings from two subjects were not included in the experiment. The speech recordings of one participant were contaminated by electrical noise (60 Hz) introduced by a faulty power adapter and could not be used for perceptual assessment. The second excluded speaker participant did not achieve resonant EL speech. The remaining four speakers were selected to produce target stimuli with the monotone and tonal EL. Speakers ranged in age from 23 to 28, with a mean age of 25.3 years. They had no reported history of speech or language problems and were able to read the given speech material in Chinese.

2.3.2. Listeners

Four native speakers of Mandarin Chinese, two male and two female, who passed a hearing screening (at 250, 500, 1000, 2000, and 4000 Hz at 25 dB HL) participated in the perceptual experiment. These individuals were not part of the Speaker participant group. Data from one subject was omitted due to poor intra-rater reliability; therefore, data analysis was performed on the three remaining listener participants, two male and one female. Listeners ranged in age from 24 to 28, with a mean age of 26.0 years. They had reported no prior exposure to EL speech.

2.4. Speaker training

Speakers received basic training for EL placement, articulation, and rate of production for both the monotone EL mode and tonal EL mode. Instructions and materials were presented on Microsoft PowerPoint slides and explained verbally. Participants were initially trained to find proper neck placement to produce the clearest-sounding EL speech. They were instructed to hold their breath while speaking to (1) prevent phonation and air flows unachievable by Laryngectomees and (2) to approximate the acoustic effects caused by anatomical changes after laryngectomy. Using the EL in a

standard, monotone fashion, each speaker practiced producing a closed set of numbers, colors, and animals in Mandarin. Verbal and tactile feedback were provided to speakers to assist with EL placement and articulation, and training concluded after 10 min.

Participants who were able to achieve adequate sound transmission to produce resonant speech then underwent a 10 min training session with the tonal EL. Using the same list of training words, speakers learned to coordinate their articulation so that it coincided with preprogrammed tonal patterns. To enable speakers to coordinate their articulation with the timing of tone output by the EL, a visual countdown of three seconds appeared on the computer screen, followed by playback of the preprogrammed tone on the tonal EL.

2.5. Stimulus materials

Target words for production consisted of 104 monosyllabic and 104 disyllabic words. These 208 words were also divided evenly into the monotone EL and tonal EL conditions. Single syllable words were randomly chosen from the China National Standards of the acoustic-speech articulation testing method (GB/T 15508-1995), which was the same word bank used in Wan et al.'s (2012) study. Disyllabic words were likewise randomly chosen, yet selected so that segmental context alone caused lexical ambiguity. Speaker participants produced one of the four sets of 26 monosyllabic and 26 disyllabic words, half with the monotone EL mode and half with the tonal EL mode (N = 13 words x 2 word lengths x 2 devices = 52 total words per speaker). Speakers were expected to say each target word only once but were given up to two chances to record their productions before advancing to the next word if errors occurred in EL triggering, unintended natural voicing, or other extraneous circumstances. Audio recordings were made with a headset microphone positioned 5 cm from the lips, and acoustic data were saved as .wav files with a sampling rate of 44,100 Hz.

2.6. Listener task and agreement

Speech samples were presented in a sound-attenuated room on headphones set to a comfortable level by the listener. All of the speech samples were combined and randomly sorted to create two sets of stimuli. The first set consisted of 208 tokens (52 words × 4 speakers) presented randomly for transcription by a custom program written in MATLAB® (The Mathworks, USA). Listeners were asked to transcribe the pinyin (phonetic transcription with corresponding tones) of the tokens on a sheet of paper, listening to each stimulus no more than twice. The second set contained 32 audio tokens (6 words × 4 speakers+8 repeated to measure intra-rater reliability), each presented along with a visual label of the target word using Microsoft PowerPoint. Listeners were asked to score the speech acceptability of each word on a 100 mm visual analog scale (VAS), ranging from 0 mm = not acceptable to 100 mm = very acceptable.

Intra-rater agreement was calculated for acceptability ratings by comparing each listener's first and second ratings of the eight repeated samples. Ratings were first converted to a seven-point equal interval scale, most commonly used for voice evaluations (Eadie and Kapsner-Smith, 2011). Two discrete judgments within ±7.14 mm on a 100 mm VAS were considered to agree exactly (Kreiman and Gerratt, 1998). One listener participant had an intra-rater agreement of only 13%, and was subsequently excluded from the study. The remaining three listeners whose data were used in the study had 50%, 75%, and 88% exact agreement, exceeding the level of chance agreement (14%). Intraclass correlation coefficients for inter-rater reliability of perceptual ratings were calculated to be .77 for intelligibility and .75 for acceptability.

3. Results

3.1. Intelligibility and perceptual accuracy

Listener transcriptions of speech samples were placed in one of four categories: correct syllable only, correct tone only, both syllable and tone correct, or neither syllable nor tone correct. An average score for categorical judgments of each stimulus was then established based on the majority response (i.e., at least two out of three listeners in agreement).

Fig. 3 depicts the percentage of listener identification for each of the four categories for words produced using the monotone EL and tonal EL. Results of a binary logistic regression analysis for intelligibility (both syllable and tone correct) indicated significant differences among stimuli based on the predictors word type, EL mode and speaker. A test of the full model with the set of predictors against the null model with no predictors was significant, $\chi 2 = 54.96$, $p < .001$; this indicates that the predictors reliably distinguish correct from incorrect samples. The approximate in-group status accounted for by the predictors was .35 using Nagelkerke's formula. Model sensitivity was 90% and specificity 63%, with an overall hit rate of 69%, when the cut value was adjusted to .173. Word type was significantly uniquely predictive of intelligibility ($b = .96$ ($SE = .38$), *Wald* (1)= 6.36, $p < .05$, $OR = 2.61$), Specifically, disyllabic words had 2.6 times higher adjusted odds [95% CI, 1.24–5.49] of being correct than monosyllabic words (controlling for EL mode). EL mode was also significantly uniquely predictive of intelligibility, $b = 2.82$ ($SE = .51$), *Wald* (1) = 30.57, $p < .001$, $OR = 16.75$; words produced using the tonal EL were 16.75 times [95% CI, 6.17–45.48] more likely to be correct than monotone words (controlling for word type, $p < .001$). Differences among speakers were not statistically significant ($p = .25$).

Table 1 shows listener response patterns for tones produced by both EL types for monosyllabic words. Tables 2 and 3 show the same information for disyllabic words. Fig. 4 compares listener identification of tones based on perceptual data for mono- and disyllabic words. Average tone identification for monosyllabic and disyllabic words were nearly identical (within 1% of each other), and were therefore combined for graphic presentation.

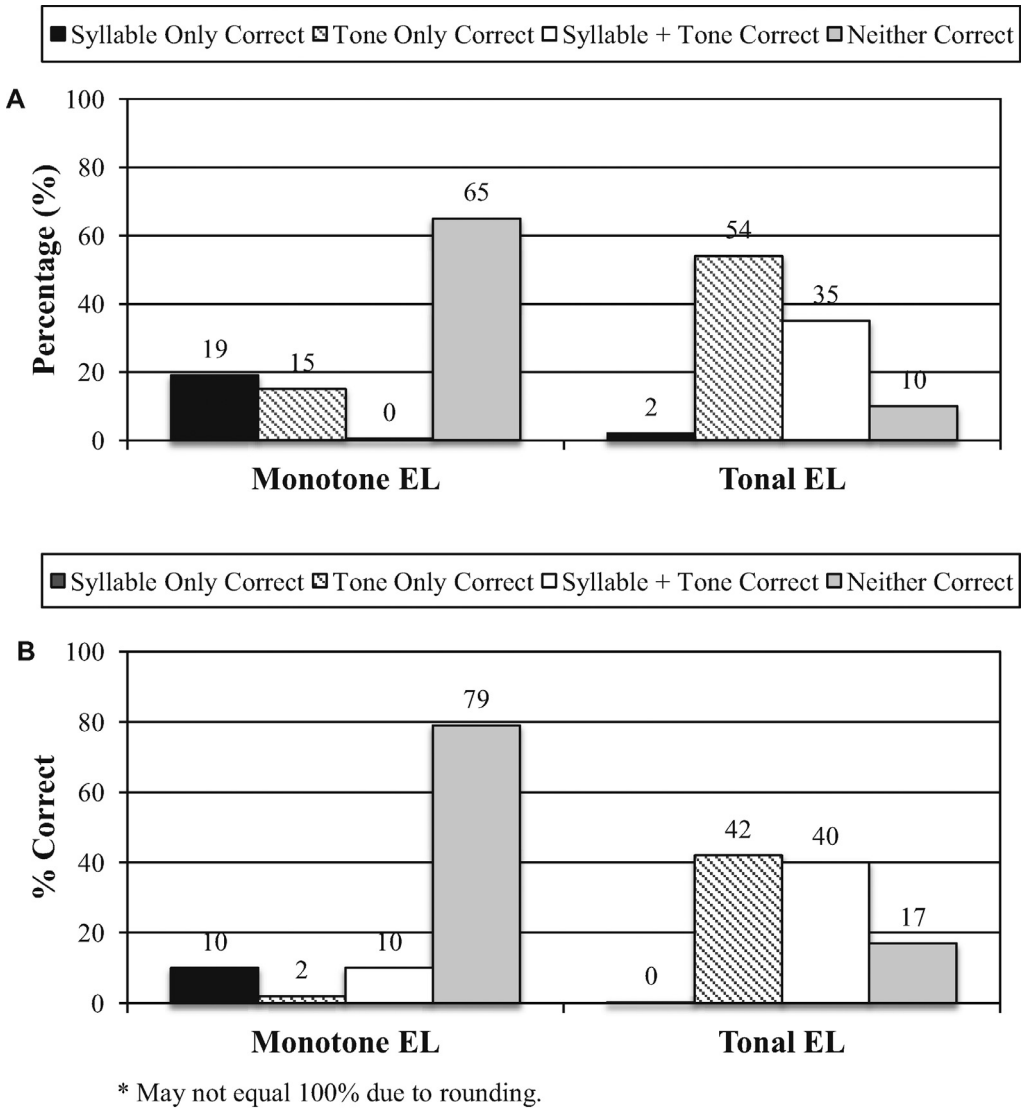

Fig. 3. Perceptual accuracy of monosyllabic words (A) and disyllabic words (B) produced with the monotone EL and tonal EL.

Table 1
Confusion matrix of responses to Mandarin tones for monosyllabic words.

Stimulus tones	Identified tones							
	Produced with monotone EL				Produced with tonal EL			
	HL	LR	FR	HF	HL	LR	FR	HF
HL	**95.8**	0.0	4.2	0.0	**100.0**	0.0	0.0	0.0
LR	90.7	**1.8**	5.6	0.0	2.4	**97.6**	0.0	0.0
FR	100.0	0.0	**0.0**	0.0	0.0	40.0	**60.0**	0.0
HF	97.4	0.0	0.0	**2.6**	0.0	0.0	0.0	**97.9**

Some rows may not equal 100% due to listeners assigning 2 tones to monosyllabic words, in which case the responses were omitted.

3.2. Acceptability

A 2 (EL mode) × 2 (word length) between-subjects ANOVA yielded a main effect for EL mode, $F(1,20) = 78.188$, $p < .001$, indicating that the mean acceptability score was significantly greater for tonal EL speech ($M = 86.4$, $SD = 10.8$) than for monotone EL speech ($M = 39.3$, $SD = 16.0$). Although the mean acceptability score was higher for disyllabic words ($M = 68.1$, $SD = 25.1$) than monosyllabic words ($M = 57.6$, $SD = 29.9$), the main effect of word length was not statistically significant ($F(1,20) = 3.888$, $p > .05$). No significant interaction was reported between EL mode and word length ($F(1,20) = .404$, $p > .05$).

4. Discussion

This study examined a method of generating tonal distinctions in Mandarin Chinese using an EL controlled by computer software. Preprogrammed tones based on previously reported average F_0 contours across speakers were created to allow for consistent tone production. To produce speech, EL users were required to coordinate computer-controlled EL voicing with movement of their articulators. Speaker participants used the customized EL to produce words in either a

Table 2
Confusion matrix of responses to Mandarin tones for disyllabic words when produced with the monotone EL.

Stimulus tones	Response tones: produced with monotone EL															
	HL +HL	HL +LR	HL +FR	HL +HF	LR +HL	LR +LR	LR +FR	LR +HF	FR +HL	FR +LR	FR +FR	FR +HF	HF +HL	HF +LR	HF +FR	HF +HF
HL+HL	**100.0**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
HL+LR	66.7	**0.0**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	33.3	0.0	0.0
HL+FR	66.7	0.0	**0.0**	0.0	16.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	16.7
HL+HF	66.7	0.0	0.0	**0.0**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	16.7	16.7	0.0	0.0	0.0
LR+HL	100.0	0.0	0.0	0.0	**0.0**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
LR+LR	66.7	0.0	0.0	0.0	0.0	**16.7**	0.0	0.0	16.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0
LR+FR	100.0	0.0	0.0	0.0	0.0	0.0	**0.0**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
LR+HF	66.7	0.0	0.0	0.0	0.0	0.0	0.0	**33.3**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
FR+HL	72.2	0.0	0.0	0.0	0.0	5.6	0.0	0.0	**16.7**	0.0	0.0	5.6	0.0	0.0	0.0	0.0
FR+LR	33.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.0**	0.0	0.0	0.0	33.3	0.0	33.3
FR+FR	66.7	0.0	0.0	6.7	6.7	0.0	**0.0**	0.0	0.0	0.0	**13.3**	6.7	0.0	0.0	0.0	0.0
FR+HF	59.3	3.7	0.0	7.4	4.0	4.0	0.0	0.0	0.0	0.0	0.0	**14.8**	0.0	0.0	4.0	4.0
HF+HL	77.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	11.1	0.0	**11.1**	0.0	0.0	0.0
HF+LR	55.6	0.0	0.0	0.0	11.1	11.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**11.1**	0.0	0.0
HF+FR	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.0**	0.0
HF+HF	76.7	0.0	0.0	3.3	0.0	0.0	3.3	0.0	0.0	6.7	0.0	0.0	0.0	0.0	0.0	**10.0**

Some rows may not equal 100% due to listeners assigning 3 tones to disyllabic words, in which case the responses were omitted.

Table 3
Confusion matrix of responses to Mandarin tones for disyllabic words when produced with the tonal EL.

Stimulus tones	Response tones: produced with tonal EL															
	HL +HL	HL +LR	HL +FR	HL +HF	LR +HL	LR +LR	LR +FR	LR +HF	FR +HL	FR +LR	FR +FR	FR +HF	HF +HL	HF +LR	HF +FR	HF +HF
HL+HL	**100.0**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
HL+LR	0.0	**83.3**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	16.7	0.0	0.0	0.0	0.0	0.0	0.0
HL+FR	0.0	0.0	**100.0**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
HL+HF	0.0	0.0	0.0	**95.2**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.0
LR+HL	0.0	0.0	0.0	0.0	**100.0**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
LR+LR	0.0	0.0	0.0	0.0	0.0	**100.0**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
LR+FR	0.0	0.0	16.7	0.0	0.0	0.0	**83.3**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
LR+HF	4.8	0.0	0.0	0.0	0.0	0.0	0.0	**85.7**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
FR+HL	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**100.0**	0.0	0.0	0.0	0.0	0.0	0.0	0.0
FR+LR	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	**n/a**	n/a	n/a	n/a	n/a	n/a	n/a
FR+FR	0.0	0.0	11.1	0.0	0.0	0.0	**11.1**	0.0	0.0	0.0	**77.8**	0.0	0.0	0.0	0.0	0.0
FR+HF	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**100.0**	0.0	0.0	0.0	0.0
HF+HL	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**83.3**	0.0	0.0	16.7
HF+LR	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**100.0**	0.0	0.0
HF+FR	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**100.0**	0.0
HF+HF	0.0	0.0	0.0	8.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	8.3	0.0	0.0	0.0	**83.3**

Some rows may not equal 100% due to listeners assigning 3 tones to disyllabic words, in which case the responses were omitted.
The FR+LR tone combination was not included in the stimuli and is labeled "n/a" as a row.

monotone mode (monotone EL) or tonal mode (tonal EL). Results indicated that adding tonal information to the EL increased the intelligibility of monosyllabic and disyllabic words by 32.5% and acceptability ratings by 47.1 out of 100 points. For words produced using the monotone EL, listeners were often unable to correctly identify either the syllable or tone of monosyllabic words (neither correct in 65% of instances) and disyllabic words (neither correct 79% of the time). These findings suggest that listeners were better able to understand and preferred speech produced with the tonal EL over the monotone EL.

Wan et al. (2012) similarly found increased word intelligibility with the addition of EL tone information via a track-ball interface (WT-EL), demonstrating the importance of tone information typically absent in EL speech. Intelligibility of single words in the present study was lower than findings reported by Wan et al. (2012) despite the use of words from the same word bank (China National Standards of the acoustic-speech articulation testing method (GB/T 15508-1995). Specifically, listeners in the present study correctly identified approximately 10% fewer words for the monotone condition and 8% fewer tonal words than in the earlier study. This may be related to differences in scoring criteria and data analysis. Whereas Wan et al. (2012) analyzed word accuracy as a continuous variable based on the average correctness of the speech samples, the current study treated intelligibility as

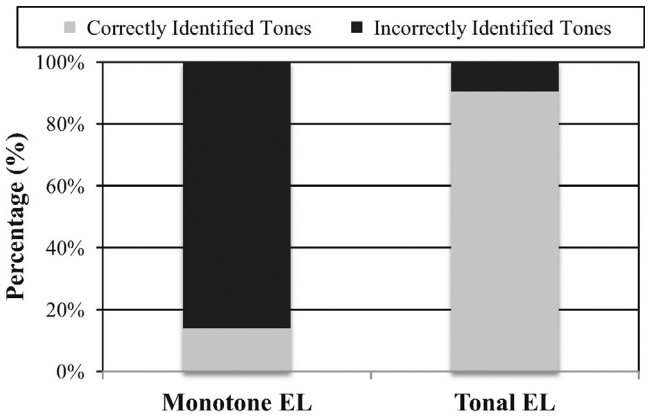

Fig. 4. The percentage of correctly and incorrectly identified tones for monosyllabic and disyllabic words (combined) produced with the monotone EL and tonal EL.

a categorical variable based on the majority correctness score (i.e. at least 2 of 3 listeners). In addition, EL speakers in the previous study were Laryngectomees who had been using a monotone EL for daily communication for over three years. Therefore, lower percentages of syllable identification in this study may be in part related to the differences in EL user proficiency. Nevertheless, overall tone identification for monosyllabic words produced using the tonal EL (89%) was comparable to that generated by the WT-EL (86%).

Patterns of tone confusion indicated differences between the tonal EL and monotone EL. As expected, most of the tones for monosyllabic words produced with the monotone EL were identified as HL. HL was identified with 95.8% accuracy, whereas LR, FR, and HF were recognized below chance level. This pattern is generally in agreement with the results reported by Wan et al. (2012), although in the current study identification of HL was slightly higher and identification of the other 3 tones was slightly lower (i.e. our listeners more consistently identified all tones as HL). For monosyllabic words produced with the tonal EL, HL, LR, and HF were correctly identified above 97%, while FR was mistaken for LR in 40% of instances.

Although tone identification was significantly higher for words produced with the tonal EL than with the monotone EL, tone identification errors suggests that at least one of the preprogrammed tone patterns used to drive the tonal EL was not entirely natural. In particular, the most common tone misidentification was labeling FR as LR (FR had only a 60% correct identification rate). Had FR been identified as accurately as the other tones, then overall tone identification for monosyllabic words would have been 98.5% correct instead of 88.9%. Interestingly, although FR was often misidentified as LR, the reverse was never true. With normal laryngeal (NL) speech, FR and LR are likewise the most commonly confused tones for monosyllabic words, but LR is perceived more often as FR than the reverse (Liu et al., 2006), which is opposite the present findings for preprogrammed EL tones. The common misidentification of FR as LR in this study suggests that the preprogrammed pattern for FR was too LR-like. If tonal confusion is related to differences between NL tonal contours and that of the tonal EL, this could be addressed by providing the tonal EL with more accurate tone patterns. For example, F_0 changes for each tone could be created by a mathematical function that not only produces naturalistic pitch patterns, but also allows the system to generate tones in relation to a wide range of habitual pitch (e.g. male versus female vocal registers) with flexible tone durations to support different speaking rates.

Additional work is needed to explore the effect of F_0 contours on tone perception, as well as the contribution of suprasegmental cues mentioned above, such as tone duration. For the present study, tone duration was kept constant to isolate the effects of F_0 contour on intelligibility and acceptability. In reality, people typically speak with consistently different durations for the four tones (Xu, 1997; Whalen & Xu, 1992). Creating preprogrammed tones that accurately reflect those typical durations may improve naturalness, with additional improvements in intelligibility and acceptability. At the same time, although tones used in this study were substantially longer than typical NL speech (450 ms per syllable; Xu, 1997), longer syllable duration is appropriate for EL speech. For example, clinical guidelines for EL speech described by Doyle (2005) recommend the use of a slightly reduced speech rate to achieve the best EL signal transmission and optimize intelligibility. The tendency for EL users to produce all tones with longer duration than NL speakers has also been documented by Liu et al. (2006). Therefore, it may be helpful to generate EL tones with increased duration for all tones, while maintaining particular duration ratios among the tones to reflect natural, relative duration patterns.

Although the present report demonstrated speakers' ability to produce intelligible mono and disyllabic words using computer-activated, preprogrammed tones, additional research is needed to determine the most practical and effortless method of selecting tones for continuous EL speech. Manual selection of individual tones could be accomplished using multiple surface-mounted buttons similar to the activation button on conventional ELs, or some other finger-controlled interface such as a multi-axis rocker switch, track ball, touch-sensitive strip, etc. Another possible tone-control interface could incorporate accelerometers or gyroscopes to detect EL movements or angles for tone selection. This may function similarly to the position-sensing capabilities of modern cameras and mobile phones, or the wrist-watch-based control of EL activation described by Matsui et al. (2013).

Regardless of the particular interface, selection of individual tones would require dedicated concentration not only on awareness of which tones are needed for each syllable (slightly in advance) throughout running speech, but also on the manual gestures to select each tone. By using the selection of preprogrammed F_0 patterns rather than the manual control of patterns we hoped to minimize the cognitive load and manual dexterity required for continuous tonal EL speech. Although fluent tone selection could be challenging at first, we hope that it might become a subconscious process in time, akin to touch-typing on a keyboard. Studies are underway

to develop an interface that allows for rapid, comfortable tone selection, and to determine Mandarin speakers' ability to select tones during running speech. In addition to young, healthy speakers like those participating in the present report, future work will need to focus on Mandarin speakers age 50 or older to better represent individuals having undergone total laryngectomy (Braithwaite, 1999), and of course test the tonal EL speaking capabilities in laryngectomized individuals versus their typical means of alaryngeal communication.

5. Conclusion

This research study was conducted with the goal of testing a method for providing tone production in Mandarin Chinese EL speech. A customized hand-held EL driven by computer software to control F_0 enabled speakers to produce monosyllabic and disyllabic words with preprogrammed F_0 variation for the four Mandarin tones (tonal EL) and without F_0 variation (monotone EL). Native Mandarin speakers with normal neck anatomy and inexperienced with EL speech were trained to articulate coincidentally with computer-controlled EL voicing (without phonating). Perceptual assessment from native Mandarin-speaking listeners indicated that words produced with the tonal EL were significantly more intelligible and acceptable than words produced with the monotone EL, and tones produced with the tonal EL were perceived more accurately. However, tones were not always identified with 100% accuracy or at the rate of NL speech, and the impact of several suprasegmental features (e.g. duration and amplitude) on tone identification has yet to be determined. Future work is needed to provide an intuitive method of tone selection and test tone selection capabilities during running speech in the Laryngectomee population.

Acknowledgments

We would like to thank Mark Robertson at Griffin Laboratories for providing the modified TruTone® EL used in the study. This project was supported by a grant from the Christopher Norman Education Fund at the MGH Institute of Health Professions and R42DC011212-02 to Mark Robertson and James Heaton, Multiple PI. We would also like to acknowledge Anthony Guarino, PhD of the Massachusetts General Hospital Institute of Health Professions for help with statistical analysis, and Jie (Kingsley) Yang, PhD for invaluable assistance as a native speaker of Mandarin.

References

Braithwaite, D.O., Thompson, T.L. (Eds.), 1999. Handbook of Communication and People with Disabilities: Research and Application. Routledge, Mahwah, NJ.

Doyle, P.C., 2005. Clinical procedures for training use of the electronic artificial larynx. In: Doyle, P., Keith, R.L. (Eds.), Contemporary Considerations in the Treatment and Rehabilitation of Head and Neck Cancer: Voice, Speech, and Swallowing. Pro-Ed; Austin, TX, pp. 545–570.

Gandour, J., Weinberg, B., Petty, S.H., Dardarananda, R., 1988. Tone in Thai alaryngeal speech. J Speech, Language, Hearing Res 53 (1), 23–29. http://doi.org/10.1044/jshd.5301.23.

Goldstein, E.A., Heaton, J.T., Kobler, J.B., Stanley, G.B., Hillman, R.E., 2004. Design and implementation of a hands-free electrolarynx device controlled by neck strap muscle electromyographic activity. IEEE Trans Biomed Eng 51 (2), 325–332. http://doi.org/10.1109/TBME.2003.820373.

Hillman, R., Walsh, M., Wolf, G., Fisher, S., Hong, W., 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. Ann Otology, Rhinology, Otolaryngology 135 (7), 704–711.

Kreiman, J., Gerratt, B.R., 1998. Validity of rating scale measures of voice quality. The J Acoustical Society America 104 (3), 1598–1608.

Liu, H., Ng, M.L., 2007. Electrolarynx in voice rehabilitation. Auris Nasus Larynx 34 (3), 327–332.

Liu, H., Wan, M., Ng, M., Wang, S., Lu, C., 2006. Tonal perceptions in normal laryngeal, esophageal, and electrolaryngeal speech of Mandarin. Folia Phoniatrica et Logopaedica 58 (5), 340–352. http://doi.org/10.1159/000094568.

Matsui, K., Kimura, K., Nakatoh, Y., Kato, Y.O., 2013. Development of electrolarynx with hands-free prosody control. In: Proceedings of the 8th ISCA, pp. 273–277.

Ng, M.L., Lerman, J.W., Gilbert, H.R., 1998. Perceptions of tonal changes in normal laryngeal, esophageal, and artificial laryngeal male Cantonese speakers. Folia Phoniatrica et Logopaedica 50 (2), 64–70.

Uemi, N., Ifukube, T., Takahashi, M., Matsushima, J., 1994, July. Paper presented at the Proceedings of the 3rd IEEE International Workshop on Robot and Human Communication, Nagoya. Design of a new electrolarynx having pitch control function. http://doi.org/10.1109/ROMAN.1994.365931.

Wan, C., Wang, E., Wu, L., Wang, S., Wan, M., 2012. Design and evaluation of an electrolarynx with tonal control function for Mandarin. Folia Phoniatrica et Logopaedica 64 (6), 290–296. http://doi.org/10.1159/000346861.

Whalen, D.H., Xu, Y., 1992. Information for Mandarin tones in the amplitude contour and in brief segments. Phonetica 49, 25–47. http://doi.org/10.1159/000261901.

Xu, Y., 1997. Contextual tonal variations in Mandarin. J Phonetics 25 (1), 62–83. http://doi.org/10.1006/jpho.1996.0034.

Zhang, J., Qi, S., Song, M., Liu, Q., 1981. On the important role of Chinese tones in speech intelligibility. Chinese J Acoustics 3, 278–283.