Published in final edited form as:
J Speech Lang Hear Res. 2009 October ; 52(5): 1360–1369. doi:10.1044/1092-4388(2009/08-0167).

Development and Perceptual Evaluation of Amplitude-Based F0 Control in Electrolarynx Speech

Yoko Saikachi
Harvard-MIT Division of Health Sciences and Technology, Cambridge, MA

Kenneth N. Stevens
Massachusetts Institute of Technology, Cambridge, MA

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

Abstract

Purpose—Current electrolarynx (EL) devices produce a mechanical speech quality that has been largely attributed to the lack of natural fundamental frequency (F0) variation. In order to improve the quality of EL speech, in the present study the authors aimed to develop and evaluate an automatic F0 control scheme, in which F0 was modulated based on variations in the root-mean-square (RMS) amplitude of the EL speech signal.

Method—Recordings of declarative sentences produced by 2 male participants before and after total laryngectomy were used to develop procedures for calculating F0 contours for EL speech. Specifically, the positive linear relationship between F0 and RMS amplitude observed in pre-laryngectomy speech was used as the basis for generating an F0 contour based on the amplitude variation of EL speech. An analysis-by-synthesis approach was used to modify the F0 contour, and a perceptual experiment was conducted to examine its impact on the quality of the EL speech.

Results—The results of perceptual experiments showed that modulating the F0 of EL speech using a linear relationship between amplitude and frequency made it significantly more natural sounding than EL speech with constant F0.

Conclusions—The current study provides preliminary support for amplitude-based control of F0 in EL speech.

An *electrolarynx* (EL) is a battery-powered device that produces a sound that can be used to acoustically excite the vocal tract as a substitute for laryngeal voice production. In the United States, the prevalence of EL use among patients is as high as 85% at 1-month postlaryngectomy (Hillman, Walsh, Wolf, Fisher, & Hong, 1998), with multiple studies reporting longer term use of an EL as a primary mode of communication by more than half of laryngectomy patients (Gray & Konrad, 1976; Hillman et al., 1998; Mendenhall et al., 2002; Morris, Smith, Van Demark, & Maves, 1992). Two ELs are currently available for use by laryngectomy patients: the neck-type (transcervical or trancutaneous) and mouth-type (transoral or intraoral). The current study focused on a neck-type device because this is by far the most commonly used type of EL.

ELs provide laryngectomy patients with the basic capability to communicate verbally (using oral speech production), provided conditions are sufficiently favorable (e.g., there is

Contact author: Robert E. Hillman, Massachusetts General Hospital Voice Center, One Bowdoin Square, (11th Floor), Boston, MA 02114. hillman.robert@mgh.harvard.edu..

minimal competing noise, the listener has normal hearing, and is familiar with EL speech). However, EL speech contains persistent acoustic deficits that result in reduced intelligibility and contribute to its “mechanical” or “nonhuman” (robotic) speech quality that often draws undesirable attention to the user. EL users have a particularly difficult time communicating with individuals who are unfamiliar with EL speech, which can make telephone use especially problematic. The main acoustic deficits associated with EL speech are (a) lack of normal fundamental frequency (F0) variation (Ma, Demirel, Espy-Wilson, & MacAuslan, 1999; Meltzner & Hillman, 2005; Uemi, Ifukube, Takahashi, & Matsushima, 1994); (b) the presence of the directly radiated signal (i.e., the buzz from the EL that is not filtered by the user's vocal tract but radiates directly to the listener; Cole, Stridharan, Moody, & Geva, 1997; Espy-Wilson, Chari, MacAuslan, Huang, & Walsh, 1998; Liu, Zhao, Wan, & Wang, 2006; Niu, Wan, Wang, & Liu, 2003; Pandey, Bhandarkar, Bachher, & Lehana, 2002; Pratapwar, Pandey, & Lehana, 2003); and (c) an improper source spectrum (Qi & Weinberg, 1991; Weiss, Yeni-Komshian, & Heinz, 1979).

Several studies have demonstrated that significant improvements in EL speech could be accomplished by adding appropriate control of F0. Some of the work has illustrated the linguistic deficits caused by a lack of F0 control (Gandour & Weinberg, 1983, 1984; Weinberg & Gandour, 1986). For example, Gandour and Weinberg (1983) conducted perceptual experiments in order to determine the degree to which EL speakers were able to achieve intonational contrasts. Results showed that users of the electrolarynx were generally unable to achieve intonational distinctions with a flat F0 contour, indicating that F0 modulation is important for the production of intonation. Lack of adequate F0 control has been shown to be even more detrimental to the intelligibility of EL users who speak tone languages such as Thai, Mandarin, and Cantonese (Gandour, Weinberg, Petty, & Dardarananda, 1988; Liu, Wan, Ng, Wang, & Lu, 2006; Ng, Gilbert, & Lerman, 2001), where F0 contours contributed most to the perception of meaning among the three main acoustic cues (F0 contour, duration, and amplitude contour; Ng et al., 2001). More recent work has examined the impact of aberrant acoustic properties on the quality of EL speech. Meltzner and Hillman (2005) demonstrated that the addition of normal F0 variation was associated with the largest improvements in the “naturalness” of EL speech, as compared with other acoustic enhancements (compensation for low-frequency deficit and reduction of noise radiated directly from the device). Ma et al. (1999) developed a postprocessing scheme in which a cepstral-based method was used to replace the original F0 contour of EL speech with a normal F0 pattern and showed that adding F0 variation clearly improved naturalness of EL speech. Although this postprocessing technique was promising, its practical applications are limited because it requires pre-recording EL speech and cannot be implemented in real time.

Adding the proper F0 variation to EL speech in real time is very challenging because it would require the means to estimate what pitch the speaker intends to use (i.e., access to underlying linguistic and/or neural processes), or utilization of alternative signals or control sources (e.g., Kakita & Hirama, 1989; Sekey & Hanson, 1982; Uemi et al., 1994). In one such approach, Uemi et al. (1994) used air pressure measurements obtained from a resistive component placed over the stoma to control the fundamental frequency of an EL, but only 2 of 16 participants were able to master control of the device. Other work has demonstrated the potential feasibility of accessing laryngeal neuromotor signals post-laryngectomy to use in controlling the onset, offset, and F0 of an EL. However, this general approach requires further testing and development and may not be effective in all EL users (Goldstein, 2003; Goldstein, Heaton, Kobler, Stanley, & Hillman, 2004; Heaton et al., 2004).

Other possibilities for controlling F0 in EL speech include implementing a fixed F0 contour (van Geel, 1982; Yahata & Ifukube, 1989) or using the frequency control supplied on some

EL devices. The impact of a fixed F0 contour on the quality of EL speech still needs to be objectively evaluated, and there is also some concern that a fixed contour could lead listeners to confuse the intent of the speaker (e.g., a question with declarative prosody). There have been attempts to include manual control of F0 in the design of some EL devices (Choi, Park, Lee, & Kim, 2001; Galyas, Branderud, & McAllister, 1982; Kikuchi & Kasuya, 2004; Takahashi, Nakao, Kikuchi, & Kaga, 2005; Tru-Tone, Griffin Laboratories, Temecula, CA), but there is considerable skepticism that manual control (e.g., pushing a button with a finger) can successfully approximate the very precise and rapid adjustments in F0 that occur during normal speech production. Furthermore, learning to effectively control F0 manually may be particularly difficult for the majority of laryngectomy patients due to their advanced age.

This brief report describes one approach that we have been developing to automatically control the F0 of EL speech. We are proposing to modulate the F0 of EL utterances based on variation in the root-mean-squared (RMS) amplitude of the EL speech signal. In previous acoustic studies of the speech of patients before (laryngeal speech) and after (EL speech) total laryngectomy, we found significant fluctuations in the amplitude of EL speech (Saikachi, Hillman, & Stevens, 2005). In particular, there was a gradual decrease of amplitude during vowels at the end of declarative utterances, which was similar to what we observed in the corresponding pre-laryngectomy speech. Furthermore, there were generally positive correlations between F0 and amplitude in pre-laryngectomy (laryngeal) speech. On the basis of these observations, we hypothesized that the amplitude variations in EL speech could be used as a basis for effectively predicting, and ultimately controlling, the F0 of EL speech in close to real time. The overall goal of this investigation was to evaluate the viability of the proposed approach by (a) developing procedures for estimating F0 on the basis of the amplitude variations in EL speech and (b) evaluating the impact of amplitude-based modulation of F0 on the quality of EL speech in perceptual experiments.

Method

Speech Recordings

In the present study, two declarative sentences from the Zoo passage[1] produced by 2 male speakers (hereafter referred to as “Speakers 1 and 2”) before and after total laryngectomy (pre-laryngectomy speech vs. EL speech) were selected from the recordings made for the Veterans Administration Cooperative Study 268 CVA-CSP 268).[2] Recording of participants from this data set who had acceptable pre-laryngectomy voice quality have been particularly useful for assessing the acoustic differences between normal (laryngeal) and EL speech and for providing acoustic “targets” to improve EL speech (Goldstein et al., 2004; Heaton et al, 2004; Meltzner, 2003; Meltzner & Hillman, 2005). Sentence 1 was “His sister Mary and his brother George went along, too.” And Sentence 2 was “You can see that they didn't have far to go.” These declarative sentences were chosen because each one terminated with vowels in which amplitude decreased consistently in both the pre-laryngectomy and EL speech of the 2 speakers (Saikachi et al., 2005).

[1]The Zoo passage used for this study is as follows: “The trip to the zoo. Last Sunday Bob went to the zoo with his mother and father. His sister Mary and his brother George went along, too. Mother packed a big basket full of good things to eat. Father took the car to the service station to get gas and have the oil checked. The family left the house at eleven o'clock and got to the zoo at twelve o'clock. You can see that they didn't have far to go.”

[2]“Hillman et al. (1998) recorded patients with advanced laryngeal cancer both before and after treatment as part of a large multi-institutional study carried out by the Cooperative Studies Program at the Veterans Administration (VA-CSP#268). Speech recordings consisting of sustained vowels, reading of a standard passage, a verbal description of a picture, and reading of a randomized list of 50 phrases (carrier phrase with different target words) were made pre- and post-treatment. Post-treatment recordings of each patient's primary means of communication (including EL speech) were made at regular follow-up visits after treatment.

The 2 speakers were chosen because they used EL speech as their primary mode of communication, the level of interference due to directly radiated EL noise was relatively low in their postlaryngectomy recordings, and their pre-laryngectomy speech was found to have relatively normal voice quality (tumor location minimally affected voice production). The 2 speakers both used a neck-placed Servox (Siemens, Munich, Germany) EL but were recorded at different VA hospitals. Of the several postlaryngectomy recordings that were made for each speaker, only the final EL speech recordings were used in this study (30 months postlaryngectomy for Speaker 1 and 12 months postlaryngectomy for Speaker 2). All recordings were made in a quiet environment using a Marantz Model 220 recorder (Marantz, Mahwah, NJ) and a Radio Shack Model 33-1071 microphone (Radio Shack Inc., Fort Worth, TX), situated 6–12 in. from the speakers (Hillman et al., 1998). An audio signal acquisition and editing software package (Syntrillium Software's Cool Edit 2000; now owned and distributed by Adobe Systems, San Jose, CA) was used to digitize the speech at 32 kHz. For this study, the speech was appropriately low-pass filtered and downsampled to 10 kHz.

Amplitude-Based F0 Estimation

Figures 1 and 2 show representative data from pre-laryngectomy and EL speech, respectively, including the audio waveform, F0 contour, and RMS amplitude as a function of time during Sentence 1. F0 was estimated using autocorrelation analysis (Markel & Gray, 1976). Both F0 and RMS amplitude were calculated every 5 ms over 40-ms intervals. Note that there is a fluctuation in amplitude over the whole utterance in both the pre-laryngectomy and EL speech. The relationship between F0 and RMS amplitude in pre-laryngectomy speech served as the basis for using the amplitude variation of EL speech to generate an F0 contour. More specifically, for each sentence and each speaker, the linear regression coefficients (intercept and slope) between F0 and amplitude were calculated for the pre-laryngectomy sentences in order to model F0 as a function of RMS amplitude. Only the voiced parts in the sentences were included for the computation. F0 values that were miscalculated by the autocorrelation methods (either halved or doubled) were also excluded from the analysis. Figure 3 shows F0 plotted against RMS amplitude for a pre-laryngectomy recording of Speaker 1 producing Sentence 1. Also shown in Figure 3 is the straight line that best fits the data, which clearly reflects the positive relationship between RMS amplitude and F0. Table 1 summarizes the regression coefficients and Pearson *r* correlation coefficients for both sentences produced by each of the 2 speakers. F0 and RMS amplitude were significantly correlated in each sentence, and the regression coefficients varied depending on the speakers and sentences.

F0 contours for the EL speech were then derived from the RMS amplitude variation in EL speech using the following equation for each sentence and speaker:

$$\text{Estimated\_ F0} = k_1 + k_2 \times \text{RMS\_ amplitude}, \quad (1)$$

where k_1 and k_2 are, respectively, the intercept and slope of the regression coefficients obtained from analyzing the pre-laryngectomy speech. Figure 4 shows an example of an amplitude-based estimate of an F0 contour superimposed on the original F0 contour for Sentence 1 produced by Speaker 1 using an EL.

Perceptual Evaluation

A perceptual experiment was conducted in order to determine (a) whether the proposed approach for controlling F0 based on amplitude could significantly improve the naturalness of EL speech and (b) whether this approach was comparable to synthesizing EL speech with an F0 contour based on pre-laryngectomy speech.

Generation of speech stimuli—The first step in generating stimuli (speech tokens) to perceptually evaluate the impact of amplitude-based F0 modulation on the quality of EL speech was to synthesize EL speech using the Klatt formant synthesizer (KLSYN). KLSYN is a well-established formant synthesizer that allows for direct control of both source and filter characteristics, and it has been shown to have the capability of producing high-quality copy synthesis for normal speech (Hanson, 1995; Klatt, 1980; Klatt & Klatt, 1990) as well as for pathological voices (Bangayan, Christopher, Alwan, Kreiman, & Gerratt, 1997). The motivation behind using this method is that synthesis can provide a tool through which the characteristics of EL speech and pre-laryngectomy speech can be compared at the level of the synthesis parameters (i.e., analysis-by-synthesis). After being parameterized, EL speech can be modified via individual or combinations of parameters to examine the resulting quality of the modified EL speech. Once copy synthesis of the original EL speech samples was accomplished, the F0 synthesis parameter was manipulated to produce EL stimuli with the desired F0 contours.

The overall scheme for generating speech tokens is shown in Figure 5. For each sentence-speaker condition, three versions of each sentence were generated from the copy-synthesized EL speech by simply modifying the F0 synthesis parameters: (a) EL speech with constant F0 (EL_S); (b) EL speech with F0 modulation based on the F0 contour of pre-laryngectomy speech (EL_f0n[3]); and (c) EL speech with F0 modulation based on the amplitude of the EL speech (EL_f0a). This resulted in 6 sentences per speaker, or a total of 12 sentences. The constant F0 values for the EL_S sentences were set to the average F0 of the pre-laryngectomy versions of the sentences, to minimize any confounding factor that could be related to differences in average F0 when comparing different stimuli. For the EL_f0a sentences, the F0 was derived from the linear relationship between F0 and amplitude in the pre-laryngectomy speech samples as described previously using Equation 1. The computed F0 was normalized such that the mean and variance of the F0 were matched to those in the pre-laryngectomy versions of the sentences. For the EL_f0n sentences, adding the pre-laryngectomy F0 contours involved two steps. First, the pre-laryngectomy and EL sentences were time aligned using the Pitch-Synchronous Overlap-Add (PSOLA) algorithm (Moulines & Charpentier, 1990), such that the phones of both sentences had the same onset times and duration. The F0 contours obtained from the time-scaled pre-laryngectomy versions of the sentences were then used to set the F0 synthesis parameters to generate the EL_f0n versions of the sentences.

Listeners—A group of 12 normal-hearing graduate students (6 females and 6 males) recruited from MIT and the Massachusetts General Hospital Institute of Health Professions served as listeners.

Experimental procedures—The synthesized stimuli were perceptually evaluated using a combination of two approaches: the Method of Paired Comparisons (PC; Meltzner & Hillman, 2005; Torgerson, 1957) and visual analog scaling (VAS).

Perceptual judgments were carried out within each of the four speaker-sentence conditions (2 Speakers × 2 Sentences = 4 conditions). Within each speaker-sentence condition, all combinations of pairs of the 3 synthesized speech tokens (3 pairs) were presented twice to listeners to total 6 paired-comparisons per condition. Thus, a total of 24 pairs of speech stimuli were presented to each listener (3 Pairs × 2 Repetitions × 4 Conditions = 24), which resulted in a total of 288 listener responses for the entire study (24 Stimulus pairs × 12 Listeners = 288).

[3]The “n” in “EL_f0n” stands for “normal.”

Before judgments were made within each of the four speaker-sentence conditions, all three speech tokens (EL_S, EL_f0n, and EL_f0a) for that condition were played to the listeners to familiarize them with the quality of the different stimuli. The pre-laryngectomy speech sample associated with the condition being evaluated was also played as a reference for the perceptual judgments. This allowed the pre-laryngectomy version of each sentence to act as an anchor so that all listeners would have a common frame of reference to make their judgments.

Each listener was seated in a sound-isolated booth and was instructed to indicate on a computer screen which of the two tokens in each pair sounded most like normal, natural speech. Then, the listener was asked to rate how different the chosen token was from normal natural speech using a VAS on the computer screen. The VAS was 10 cm long, with the left end labeled “Not At All Different” and the right end labeled “Very Different.” The presentation order of four speaker-sentence conditions was randomized for each listener. Participants were allowed to listen to the pre-laryngectomy speech token associated with each condition (anchor) as often as they wanted during both PC and VAS components of the experiment.

Data analysis—The PC data were first analyzed by conducting binomial testing in order to test the significance of the results. The PC data were also converted to scale rankings using Thurstone's Law of Comparative Judgment (Thurstone, 1927), in which speech tokens that were most consistently judged to sound more like normal, natural speech across all listeners were given a higher ranking (Meltzner & Hillman, 2005). The data from the VAS procedure were analyzed by computing the distance (in cm) from the left end of the VAS. These distances were used to calculate a mean distance for each speech type and taken as an estimate of how different a listener judged a speech token to be from natural, normal speech.

Results

The reliability of listeners was evaluated by calculating the percentages of agreement in preference judgments made by each listener in response to the repeated presentation of all token pairs. The average intralistener agreement across all four speaker-sentence conditions (speaker1-sentence1, speaker1-sentence2, speaker2-sentence1, and speaker2-sentence2) for the PC task was 80.0 ± 16.1% (range = 50%–100%), 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 .83 ± .16 (range = .52–.99).

The PC response data are summarized in Table 2. Shown are the total number and percentage of times that listeners judged each of the three different speech tokens to sound more normal or natural than the other two tokens in paired comparisons. The binomial test showed that there was a significant overall preference by the listeners for the F0-modulated EL speech (EL_f0a and EL_f0n tokens) as compared with the EL speech having constant F0 (EL_S tokens; $p < .01$). The exception was the EL_f0a vs. EL_S token pair for Sentence 2 produced by Speaker 1. Conversely, there was no significant preference for the EL_f0n tokens over EL_f0a tokens.

A summary of the overall results obtained using the PC and VAS procedures across all four speaker-sentence conditions is shown in Table 3. Note that speech types judged to be closer to normal speech received higher PC scale values and lower VAS values. The rankings of the speech types by the two scaling procedures were identical. EL speech with amplitude-modulated F0 (EL_f0a) was judged to sound better than EL speech with constant F0 (EL_S)

but not quite as good as EL speech produced with the pre-laryngectomy F0 contour (EL_f0n).

Discussion

In this study, an approach for amplitude-based control of F0 in EL speech was developed, and its impact on the quality of EL speech was examined. The approach used the positive linear relationship that was observed between F0 and amplitude in the pre-laryngectomy speech of EL users. The results of both PC and VAS experiments demonstrated that EL speech with amplitude-based F0 modulation was judged to sound more natural than EL speech with constant F0, thus lending preliminary support for using this simple linear relationship to compute an F0 contour for EL speech. Furthermore, analysis of the PC data using the binomial testing showed that there was no significant preference for the pre-laryngectomy F0 contour over amplitude-based F0 modulation, implying that the listeners found these two types of stimuli relatively similar to each other. The scale values computed by analyzing the PC data also indicate that the perceptual distance between these two types of stimuli was relatively small. Compared with previously implemented F0 control methods using a finger-controlled button (Choi et al., 2001; Galyas et al., 1982; Kikuchi & Kasuya, 2004; Takahashi et al., 2005; Tru-Tone) or stoma air-pressure measurements (Sekey & Hanson, 1982; Uemi et al., 1994), the proposed F0 control scheme does not require access to alternative signals or control sources and may not require extensive experience or training. Furthermore, this approach has the potential to be implemented with relative ease in close to real time using a prototype (portable) digital signal processor (DSP)–based hardware platform. One possible configuration could entail using the DSP system to estimate the RMS amplitude of EL speech from a microphone signal and to then generate an F0 contour (based on linear prediction) that could be fed back to drive the EL device in a real-time loop. We have already had some initial success in creating a proof-of-concept DSP-based system that operates in this fashion.

It must be noted, however, that this study was restricted to the improvement of the naturalness of declarative sentences. As described in the introduction, the F0 contour is important not only for the perceived naturalness of the EL speech but also for communicating linguistic contrasts such as intonation (e.g., declarative vs. interrogative) and contrastive stress. For example, interrogative sentences are associated with a maximal rise in F0 at the terminal portion of the utterance, whereas declarative versions are associated with a fall in the F0 during the terminal portion (Atkinson, 1976). It has been also shown that in stress-accent languages, such as American English and Dutch, stress and accent are separate linguistic constructs, and both have unique phonetic correlates (Okobi, 2006; Sluijter, 1995; Sluijter, van Heuven, & Pacilly, 1997). More specifically, in these languages, F0 movement and overall intensity are acoustic correlates of pitch accents but not of stress, which is characterized by the longer duration and high-frequency emphasis. We are in the process of conducting additional studies that include sentences specially designed to vary intonation, pitch accents, and stress patterns in order to determine the capabilities of amplitude-based control of F0 in different prosodic contexts.

Another area of inquiry has to do with the potential source of amplitude fluctuation in EL speech. One potential source of the amplitude variation in the EL speech is changes in mouth opening during speech production, but this did not always seem to account for the magnitude of the observed variation. It is possible that the user manipulates the pressure of the EL device against the neck, in a manner similar to a body or hand gesture that occurs during speech production. This manipulation could influence the pressure against the neck and, therefore, modify the amplitude of the acoustic source that excites the vocal tract. Another possibility is that the low-frequency deficit of the EL device decreases the first

formant amplitude of high vowels more than low vowels, so there is a vowel-dependent fluctuation in amplitude. We are in the process of conducting additional studies of pre-recorded EL speech (in digital audio and video format) from patients with laryngectomees (Goldstein, 2003) to evaluate hypothesized changes in amplitude due to movement of formant frequencies, changes in formant bandwidths, the degree of low-frequency deficit, and the degree of mouth opening. To examine the potential role of EL location and contact pressure, new recordings of laryngectomy EL users are being made using video recordings and a sensor on the head of the EL to measure the pressure exerted against the neck. A clearer understanding of the sources could potentially lead to improved algorithms for real-time enhancement of EL speech based on processing of the EL speech output. It could also suggest ways of training an EL user to manipulate the device to produce more natural prosody.

As the results of the VAS revealed, the rating for the best token, EL_f0n, was to the right of the midpoint of the scale (toward the “very different” end), suggesting that there were still other important acoustic factors that need to be addressed to improve the quality of the EL speech in addition to F0 modulation. This finding is consistent with the previous studies on the enhancement of the EL speech (Melztner, 2003; Meltzner & Hillman, 2005). Other important acoustic properties include deficits due to the acoustic characteristics of the EL voicing source and its location away from the terminal end of the vocal tract (i.e., introduction of spectral zeroes into the speech output) and additional modifications in the vocal tract transfer function due to the impact of the laryngectomy operation on the upper airway (Meltzner, 2003; Myrick & Yantorno, 1993). The analysis-by-synthesis approach developed in this study using KLSYN should provide the means for investigating (via generating stimuli for perceptual experiments) and testing attempts to correct (via modifying synthesis parameters) additional acoustic deficits in EL speech.

Another possible future addition to this work is to examine the implications for the laryngecomized patients who are native speakers of tone languages. As reviewed in the introduction, a lack of adequate F0 control has largely limited the ability of the EL users to signal tonal contrasts (Gandour et al., 1988; Liu, Wan, et al., 2006; Ng et al., 2001). In this context, it is interesting to note that in Mandarin Chinese, amplitude has been suggested to contribute to tone recognition when F0 information was removed (Fu & Zeng, 2000; Liu & Samuel, 2004; Whalen & Xu, 1992). It has been further demonstrated that this amplitude-based tone recognition was directly related to the correlation between amplitude contour and F0 contour, indicating that participants might have interpreted amplitude changes as F0 changed (Fu & Zeng, 2000). More research on the acoustic characteristics of tone languages in EL speech might be needed to extend the scope of our study for tone languages.

Although this investigation demonstrated preliminary feasibility of the amplitude-based F0 control of an EL, it was meant to essentially demonstrate a proof-of-concept and was, therefore, limited with respect to number of participants, sentences, and stimuli used in the perceptual experiments. Thus, the generalizability of the present results must be viewed with caution. We also did not test whether just any variation in EL F0 that is not linked to amplitude would also produce a similar level of preference when compared with a lack of F0 modulation. More research is needed to address these limitations.

Acknowledgments

This research was supported by National Institutes of Health Grants R01 DC006449 and R41 DC008722-01A1. We would like to thank Harold Cheyne, Jam es Heaton, Geoff Meltzner, and members of the Speech Communication group for their help with this experiment as well as for their comments and discussion.

References

Atkinson JE. Inter- and intraspeaker variability in fundamental voice frequency. The Journal of the Acoustical Society of America. 1976; 60:440–445. [PubMed: 993467]

Bangayan P, Christopher L, Alwan AA, Kreiman J, Gerratt BR. Analysis by synthesis of pathological voices using the Klatt synthesizer. Speech Communication. 1997; 22:343–368.

Choi HS, Park YJ, Lee SM, Kim KM. Functional characteristics of a new electrolarynx “Evada” having a force sensing resistor sensor. Journal of Voice. 2001; 15:592–599. [PubMed: 11792038]

Cole D, Stridharan S, Moody M, Geva S. Application of noise reduction techniques for alaryngeal speech enhancement. Proceedings of IEEE Region 19 Annual Conference Speech and Image Technologies for Computing and Telecommunications. 1997; 2:491–494.

Espy-Wilson CY, Chari VR, MacAuslan JM, Huang CB, Walsh MJ. Enhancement of electrolaryngeal speech by adaptive filtering. Journal of Speech, Language, and Hearing Research. 1998; 4:1253–1264.

Fu Q-J, Zeng FG. Identification of temporal envelope cues in Chinese tone recognition. Asia Pacific Journal of Speech, Language, and Hearing. 2000; 5:45–57.

Galyas, K.; Branderud, P.; McAllister, R. The “intonator.” Development of an electrolarynx with intonation control. In: Sekey, A., editor. Electroacoustic analysis and enhancement of alaryngeal speech. Charles C Thomas; Springfield. IL: 1982. p. 184-189.

Gandour J, Weinberg B. Perception of intonational contrasts in alaryngeal speech. Journal of Speech and Hearing Research. 1983; 26:142–148. [PubMed: 6865370]

Gandour J, Weinberg B. Production of intonation and contrastive stress in electrolaryngeal speech. Journal of Speech and Hearing Research. 1984; 27:605–612. [PubMed: 6521468]

Gandour J, Weinberg B, Petty SH, Dardarananda R. Tone in Thai alaryngeal speech. Journal of Speech and Hearing Disorders. 1988; 53:23–29. [PubMed: 3339865]

Goldstein, EA. Unpublished doctoral dissertation. Harvard University; 2003. Prosthetic voice controlled by muscle electromyographic signals.

Goldstein EA, Heaton JT, Kobler JB, Stanley GB, Hillman RE. Design and implementation of a hands-free electrolarynx device controlled by neck strap muscle electromyographic activity. IEEE Transactions on Biomedical Engineering. 2004; 51:325–332. [PubMed: 14765705]

Gray S, Konrad HR. Laryngectomy: Postsurgical rehabilitation of communication. Archives of Physical Medicine and Rehabilitation. 1976; 57:140–142. [PubMed: 1267585]

Hanson H. Synthesis of female speech using the Klatt formant synthesizer. MIT Speech Communication Group Working Papers. 1995; 10:84–103.

Heaton JT, Goldstein EA, Kobler JB, Zeitels S, Randolph G, Walsh M, et al. 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. 2004; 109:972–980.

Hillman RE, Walsh MJ, Wolf GT, Fisher SG, Hong WK. 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. Annals of Otology, Rhinology and Laryngology. 1998; 172(Suppl.):1–27.

Kakita Y, Hirama J. Controls of prosodic information and voiceless consonants for the electronic larynx. Technical Report of IECE. 1989; SP88-148:25–30.

Kikuchi Y, Kasuya H. Development and evaluation of pitch adjustable electrolarynx. Speech Prosody. 2004; 2004:761–764.

Klatt DH. Software for a cascade/parallel formant synthesizer. The Journal of the Acoustical Society of America. 1980; 67:971–995.

Klatt DH, Klatt LC. Analysis, synthesis, and perception of voice quality variations among female and male talkers. The Journal of the Acoustical Society of America. 1990; 87:820–857. [PubMed: 2137837]

Kreiman J, Gerratt BR, Kempster GB, Erman A, Berke G. Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research. Journal of Speech and Hearing Research. 1993; 36:21–40. [PubMed: 8450660]

Liu S, Samuel AG. Perception of Mandarin lexical tones when F0 information is neutralized. Language and Speech. 2004; 47:109–138. [PubMed: 15581188]

Liu H, Wan M, Ng LM, Wang S, Lu C. Tonal perceptions in normal laryngeal, esophageal, and electrolaryngeal speech of Mandarin. Folia Phoniatrica et Logopedica. 2006; 58:340–352.

Liu H, Zhao Q, Wan M, Wang S. Application of spectral subtraction method on enhancement of electrolarvnx speech. The Journal of the Acoustical Society of America. 2006; 120:398–406. [PubMed: 16875235]

Ma, K.; Demirel, P.; Espy-Wilson, C.; MacAuslan, J. Improvement of Electrolarvnx speech by introducing normal excitation information. Proceedings of the European Conference on Speech Communication and Technology; Budapest. 1999. p. 323-326.

Markel, JD.; Gray, AH. Linear prediction of speech. Springer-Verlag; New York: 1976.

Meltzner GS. Perceptual and acoustic impacts of aberrant properties of electrolaryngeal speech. Dissertation Abstracts International: Section B. Sciences and Engineering. 2003; 64(09):41–62.

Meltzner GS, Hillman RE. Impact of aberrant acoustic properties on the perception of sound quality in electrolarynx speech. Journal of Speech, Language, and Hearing Research. 2005; 48:766–779.

Mendenhall WM, Morris CG, Stringer SP, Amdur RJ, Hinerman RW, Villaret DB, Robbins KT. Voice rehabilitation after total laryngectomy and postoperative radiation therapy. Journal of Clinical Oncology. 2002; 20:2500–2505. [PubMed: 12011128]

Morris HL, Smith AE, Van Demark DR, Maves MD. Communication status following laryngectomy: The Iowa experience 1984–1987. Annals of Otology, Rhinology and Laryngology. 1992; 101:503–510.

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

Myrick, R.; Yantorno, R. Vocal tract modeling as related to the use of an artificial larynx. Bioengineering Conference Proceedings of the 1993 IEEE Nineteenth Annual Northeast; Mar. 1993 p. 75-77.

Ng ML, Gilbert HR, Lerman JW. Fundamental frequency, intensity, and vowel duration characteristics related to perception of Cantonese alaryngeal speech. Folia Phoniatrics et Logopaedica. 2001; 53:36–47.

Niu HJ, Wan MX, Wang SP, Liu HJ. Enhancement of electrolarynx speech using adaptive noise cancelling based on independent component analysis. Medical and Biological Engineering & Computing. 2003; 41:670–678. [PubMed: 14686593]

Okobi, AO. Unpublished doctoral dissertation. MIT; 2006. Acoustic correlates of word stress in American English.

Pandey, PC.; Bhandarkar, SM.; Bachher, GK.; Lehana, PK. Enhancement of alaryngeal speech using spectral subtraction. IEEE Digital Signal Processing Workshop 2002; 2002. p. 591-594.

Pratapwar, SS.; Pandey, PC.; Lehana, PK. Reduction of background noise in alaryngeal speech using spectral subtraction with quantile based noise estimation. Seventh World Multiconference on Systemics, Cybernetics and Informatics; 2003. p. 408-413.

Qi YY, Weinberg B. Low-frequency energy deficit in electrolaryngeal speech. Journal of Speech and Hearing Research. 1991; 34:1250–1256. [PubMed: 1787706]

Saikachi Y, Hillman RE, Stevens KN. Analysis by synthesis of Electrolarynx speech. The Journal of the Acoustical Society of America. 2005; 118:1965.

Sekey, A.; Hanson, R. Laryngectomee speech support system with prosodic control. In: Sekey, A., editor. Electroacoustic analysis and enhancement of alaryngeal speech. Charles C Thomas; Springfield, IL: 1982. p. 166-183.

Sluijter, AMC. Unpublished doctoral dissertation. Holland Institute of Generative Linguistics; 1995. Phonetic correlates of stress and accent.

Sluijter AMC, van Heuven VJ, Pacilly JJA. Spectral balance as a cue in the perception of linguistic stress. The Journal of the Acoustical Society of America. 1997; 101:503–513. [PubMed: 9000741]

Takahashi H, Nakao M, Kikuchi Y, Kaga K. Alaryngeal speech aid using an intra-oral electrolarynx and a miniature fingertip switch. Auris Nasus Larynx. 2005; 32:157–162. [PubMed: 15917173]

Thurstone LL. A law of comparative judgment. Psychology Review. 1927; 34:273–286.

Torgerson, WS. Theory and methods of scaling. Wiley; New York: 1957.

Uemi, N.; Ifukube, T.; Takahashi, M.; Matsushima, J. Design of a new electrolarynx having a pitch control function. IEEE Workshop on Robot and Human; 1994. p. 198-202.

van Geel, RC. Semi-automatic pitch control for an electrolarvnx. In: Sekey, A., editor. Electroacoustic analysis and enhancement of alaryngeal speech. Charles C Thomas; Springfield, IL: 1982. p. 190-197.

Weiss MS, Yeni-Komshian GH, Heinz JM. Acoustical and perceptual characteristics of speech produced with an electronic artificial larynx. The Journal of the Acoustical Society of America. 1979; 65:1298–1308. [PubMed: 458051]

Weinberg B, Gandour J. Prosody in alaryngeal speech. Seminars in Speech and Language. 1986; 7:95–107.

Whalen DH, Xu Y. Information for Mandarin tones in the amplitude contour and in brief segments. Phonetica. 1992; 49:25–47. [PubMed: 1603839]

Yahata H, Ifukube T. Electrolarynx providing voice pitch pattern. Japanese Journal of Logopedics and Phoniatrics. 1989; 30:309–315.

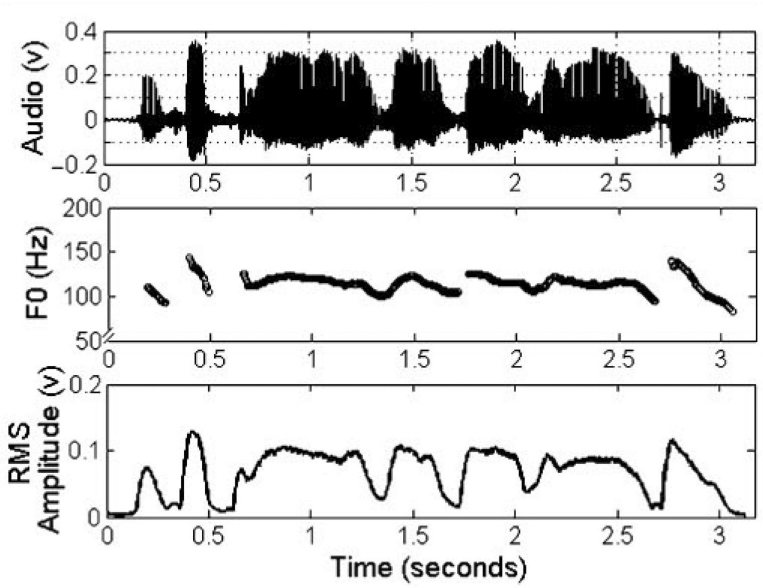

Figure 1.

Audio waveforms, F0, and RMS amplitude over time for sentence 1, “His sister Mary and his brother George went along, too/” recorded before laryngectomy (pre-laryngectomy laryngeal speech) by speaker 1. F0 = fundamental frequency; RMS = root-mean-squared; Hz = Hertz; v = volts.

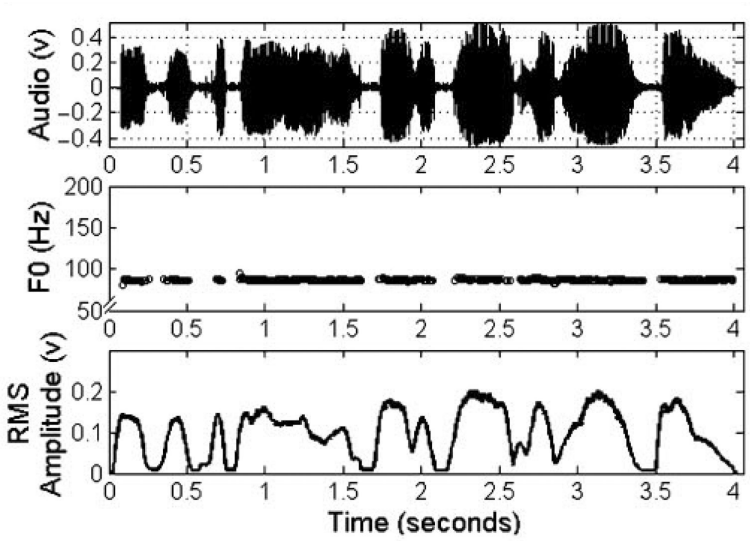

Figure 2.
Audio waveforms, F0, and RMS amplitude, and F0 over time for sentence 1, recorded by speaker 1 using an electrolarynx (EL) after laryngectomy.

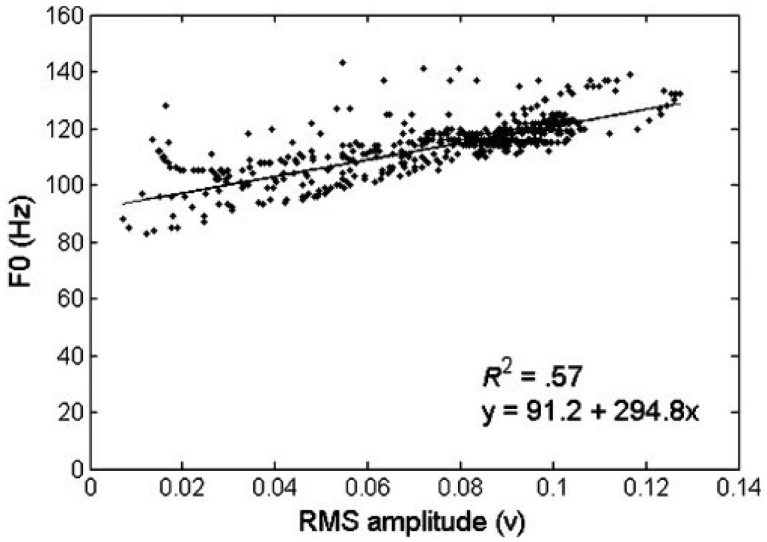

Figure 3.
F0 versus RMS amplitude and linear regression for sentence 1 produced by speaker 1. Correlation coefficients and regression coefficients are shown at the bottom.

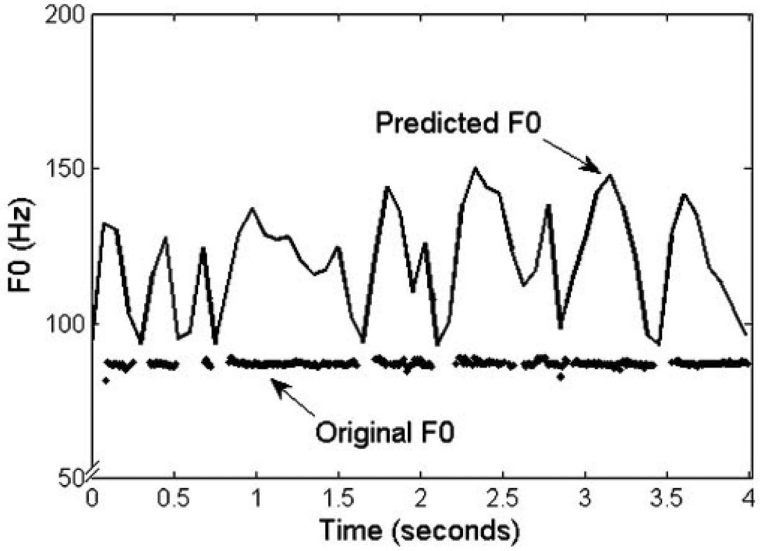

Figure 4.
Measured original F0 and amplitude-based estimates of F0 as a function of time for sentence 1 produced by speaker 1.

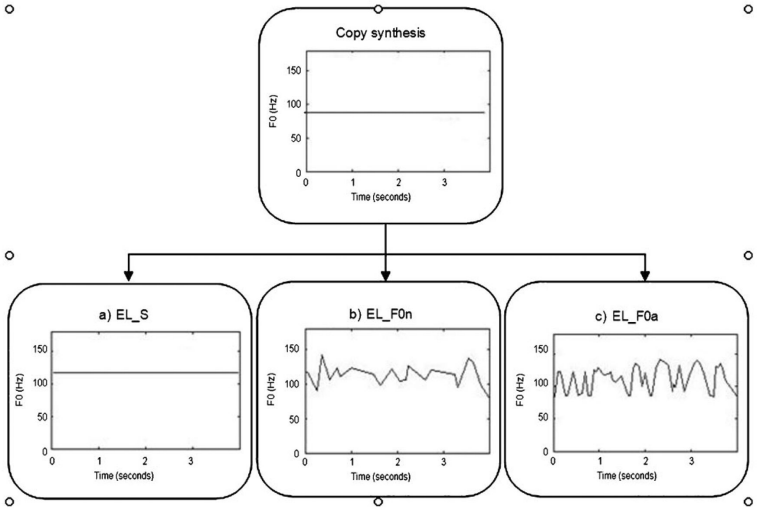

Figure 5.

F0 synthesis contours for sentence 1 and speaker 1 that were used to generate the EL speech stimuli for the perceptual experiments. “EL_S” corresponds to copy-synthesized EL speech with constant F0. “EL_f0n” and “EL_f0a” are the EL speech with F0 modulations based on the pre-laryngectomy F0 contour and EL speech amplitude, respectively.

Table 1

Values of intercept, slopes, and correlation coefficients for the different speakers and sentences.

	Speaker 1		Speaker 2	
Variable	**Sentence 1**	**Sentence 2**	**Sentence 1**	**Sentence 2**
Intercept (Hz)	91.2	102.6	92.0	91.4
Slope (Hz/volts)	294.8	262.4	190.5	182.2
Correlation coefficients (R^2)	.57*	.44*	.39*	.38*

*$p < .001$.

Table 2

Number and percentage of responses showing preference for the first token listed in each paired comparison (PC).

	Speaker 1		Speaker 2		
PC	**Sentence 1**	**Sentence 2**	**Sentence 1**	**Sentence 2**	**Overall**
EL_f0a vs. EL_S	95.8%	70.8%	95.8%	95.8%	89.6%
	23/24*	17/24	23/24*	23/24*	86/96*
EL_f0n vs. EL_S	88.0%	91.0%	92.0%	100.0%	96.0%
	21/24*	20/22*	22/24*	23/23*	89/93*
EL_f0n vs. EL_f0a	56.5%	62.5%	29.2%	75.0%	55.8%
	13/23	15/24	7/24	18/24	53/93

Note. EL_f0a = EL speech with F0 modulation based on the amplitude of the EL speech; EL_S = EL speech with constant F0; EL_F0n = EL speech with F0 modulation based on the F0 contour of pre-laryngectomy speech.

* $p < .01$.

Table 3

Overall paired comparison (PC) and visual analog scale (VAS) values.

	PC		VAS			
Speech type	**Rank**	**Scale value**	**Rank**	**Rating**	***SE***	***N***
EL_f0n	1	1.63	1	6.5	0.17	117
EL_f0a	2	1.37	2	6.9	0.18	107
EL_S	3	0.0	3	7.3	0.09	13

Note. *SE* = standard error.