# Assessment of a Method for the Automatic On/Off Control of an Electrolarynx via Lip Deformation

**Congying Wan, Liang Wu, Huixiong Wu, Supin Wang, and Mingxi Wan,** *Xi' an, P. R. China*

**Summary: Objectives.** To reduce the inconvenience caused by operating a button-electrolarynx (EL) by hand, we proposed a method for the automatic control of an EL via lip deformation and applied the method to a video-based experimental system (video-EL). The purpose of this study was to validate the method and assess its performance in producing Mandarin Chinese.

**Study Design and Methods.** Eight subjects, including one laryngectomee, were invited to participate in the assessment. First, the empirical parameters of phonation onset/offset estimation were compared with the optimal parameters obtained by minimizing simulation errors during video-EL. Second, a reaction time test was used to evaluate the ability of subjects to pronounce a single word with video-EL. Third, the fluency of subjects in producing long sentences with video-EL was calculated. Finally, the intelligibility of speech produced with video-EL was compared with that produced with button-EL.

**Results.** The empirical parameters were not significantly different from the optimal parameter and resulted in fewer interruptions during voicing. Video-EL resulted in slower voice initiation and termination when compared with button-EL, which affected the intelligibility of an isolated word. However, video-EL provided a sufficiently fluent voice source so that the intelligibility of speech produced with video-EL was not significantly different from speech produced with button-EL when producing sentences.

**Conclusions.** The method proposed in this study is effective in the automatic on/off control of an EL. Subjects produced fluent speech with video-EL that was as intelligible as that produced with button-EL when Mandarin sentences were produced continuously.

**Key Words:** Electrolarynx–Phonation onset/offset–Automatic control–Electrolaryngeal speech.

## INTRODUCTION

Laryngectomy, which is an important treatment for larynx cancer and trauma, leads to inability of speaking for the patients because of the removal of entire larynx. Because the articulators are still functional, prosthetic devices can be used to produce substituted voice source. Therefore, electrolarynx (EL), which is an external vibration-generating device, is used for exciting vocal track to produce substituted voiced source. EL is effective for communication in many situations because it is easy to use and can produce long sentence without special care.[1,2] Therefore, EL is the most commonly used prosthesis for voice rehabilitation.[3]

However, it is not very convenient to use a conventional button-EL, especially in hands-requiring situation. It is because that the on/off of conventional EL is controlled by a button switch, and it occupies one hand to hold and control it during speaking. The inconvenience brought by one hand requiring is noted in the top five deficits of EL communication for users.[4] EL devices with more convenient control manner can benefit EL users a lot.

To improve the convenience of EL, some researchers made effort to provide hand-free vibrator and more convenient switch. Zwitman et al[5,6] developed a denture-based intraoral EL, which was turned on/off by tongue. Takahashi et al[7] used a wireless miniature fingertip switch to control the denture-based vibrator. The fingertip switch had two buttons for providing command of voicing and accent via infrared communication. Hashiba et al[8] also applied a wireless controller to a wearable EL. The neckwear vibrator was embedded on a thermoplastic brace attaching to neck, the switch was bound on forefinger, and the on/off order was transmitted to vibrator controller by a wireless emitter.

Although these improvements can release users' hands from holding the device, they still need to control the switch manually. To trigger the vibrator automatically, electromyography (EMG) of reserved articulators was used. Heaton et al[9,10] found that the neck surface EMG (sEMG) signal was appropriate for obtaining phonation-related activity after transferring recurrent laryngeal nerves to denervated neck strap muscles. A hands-free EL device controlled by neck strap muscle EMG activity was developed.[11] sEMG signal was processed to obtain a slow envelope for pitch modulation and a fast envelope for initial/terminal control. Users could produce articulatory coordinated voice onset, maintenance, offset, and semantically appropriated pitch modulation with the device after training.[12,13] However, the requirement of certain nerve operation was a restriction of EMG-EL. Stepp et al[14] investigated the performance of EMG-EL controlled by EMG activities recorded from several locations. He concluded that the superior ventral neck or submental surface could provide at least one of the two best control locations, and special effort to preserve musculature for EMG-EL control was not necessary.

The EMG-EL has been proven effective for hands-free EL control. However, the electrode has to be attached on the surface of the skin, which may be uncomfortable for some users.

Accepted for publication March 16, 2012.

From the The Key Laboratory of Biomedical Information Engineering of Ministry of Education, Department of Biomedical Engineering, School of Life Science and Technology, Xi' an Jiaotong University, Xi' an 710049, P. R. China.

Address correspondence and reprint requests to Supin Wang and Mingxi Wan, The Key Laboratory of Biomedical Information Engineering of Ministry of Education, Xi' an Jiaotong University, Xi' an 710049, P. R. China. E-mail: spwang@mail.xjtu.edu.cn and mxwan@mail.xjtu.edu.cn

Journal of Voice, Vol. 26, No. 5, pp. 674.e21-674.e30
0892-1997/$36.00

doi:10.1016/j.jvoice.2012.03.002

In addition, fixing the electrode on the skin can be problematic during long periods because of neck or facial movements. Noncontact phonation onset/offset information acquisition avoids such problems. In this study, lip movement was chosen as a means of automatic control of an EL for two reasons: first, lip movement occurs in most phonation except when pronouncing particular consonants, and second, lip movement can be captured without contact using a video camera. Accordingly, an automatic EL control method based on real-time video processing is proposed. In this method, lip movement is captured with a camera, the outer contour of the lip is extracted, and the change in the outer contour of the lip is taken as a measure of lip deformation. Phonation onset/offset is then obtained from this measurement of lip deformation. The method was implemented in a video-based experimental system (termed video-EL). The system included a camera for capturing lip movement and a computer for calculating lip deformation in real time and generating automatic on/off commands to a wearable EL. To validate this method, several aspects of the system were assessed, including the parameter settings, switching performance, and speech intelligibility. The results indicate that this method is effective for the automatic control of an EL vibrator.

## METHODS

### Method for the automatic control of an EL

The proposed method for automatic control of an EL via lip deformation is shown in Figure 1. First, lip image sequences are captured with a camera and then transmitted to a computer processor. The lip area is then tracked and processed in real time. A shape feature is used to represent lip deformation. The onset/offset of phonation is thus obtained by monitoring the degree of lip deformation. Finally, an on/off command is generated via a controller to a wearable EL vibrator.

**Real-time lip deformation detection.** Several sophisticated methods for the detection of lip deformation have been proposed in the field of audio-visual speech recognition. The deformable template method[15,16] has been adopted and simplified for real-time processing. The procedure, shown in Figure 2, is introduced as follows:

*Preprocessing.* To segment the lip area accurately, a chromatic filter based on lip chromatic analysis[17] is applied to obtain an enhanced image with a high contrast between the skin and lips. The expression for the lip chromatic filter $Z$ is shown below in Equation 1.

$$Z = 0.493R - 0.589G + 0.026B \quad (1)$$

The filter Z is then applied to the region of interest (ROI) to obtain an enhanced gray image for subsequent segmentation.

*Segmentation.* The Otsu method[18] is used for its simplicity and stability to segment the lip area from the skin. A threshold $T$ divides the pixels into two classes, skin and lip. The optimal $T$ is achieved by minimizing the within-class variance, which is the same as maximizing the between-class variance.[18] The lip area is shown in a binary image by applying the optimal $T$ to the enhanced gray image.

*Template Fitting.* A deformable template is used to represent the shape of the lip outer contour. For simplicity, the asymmetric part and the orientation of the lips are ignored. The lip outer contour is fitted into an ellipse, which is determined by two parameters, the semimajor ($a$) and semiminor ($b$) axes. The center point of the lips is set as the coordinate origin, and a Hough transform is used to obtain the parameters ($a$ and $b$) of the ellipse. The parameters of the ellipse are then used to characterize the lip shape.

*Optimization of Real-Time Processing.* In the classic Otsu method and Hough transform, the optimal parameters are searched for all possible values, which is very time consuming. In the implemented system, the sampling frame is 20 frames per second (FPS). Considering that the peak velocity of lip movement is about 130 mm/s during normal speech,[19,20] the lip can only move approximately 6.5 mm at most during one sampling interval equaling 50 ms. The change between two subsequent images is therefore not very large; thus, the parameters of image N can be used to narrow the search range of the subsequent image (N + 1). A pilot study on parameter calculation showed that this range-narrowing method could achieve optimal parameters as long as an appropriate range was selected. Accordingly, the ROI of image (N + 1) is determined by $a_n$ and $b_n$, and the parameters in image (N + 1) ($T_n$, $a_n$, and $b_n$) are searched for in a range of parameters in image N.

**EL on/off control.** The lip shape is represented by an ellipse with parameters $a$ and $b$. The variation in the ratio of $b$ to $a$ ($b/a$)

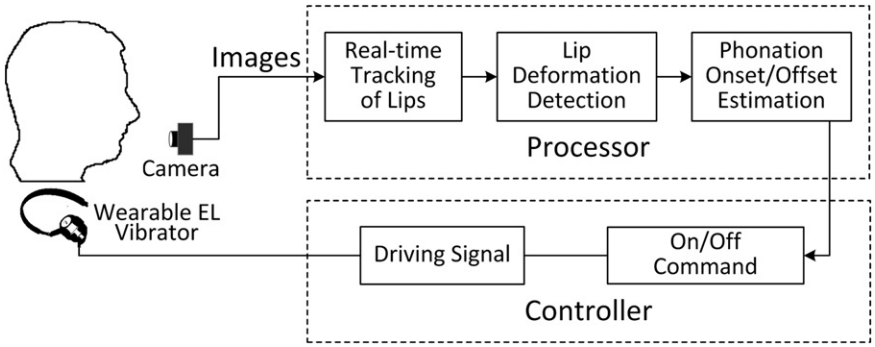


**FIGURE 1.** Schematic illustrating the process for automatic control of an EL via lip deformation.

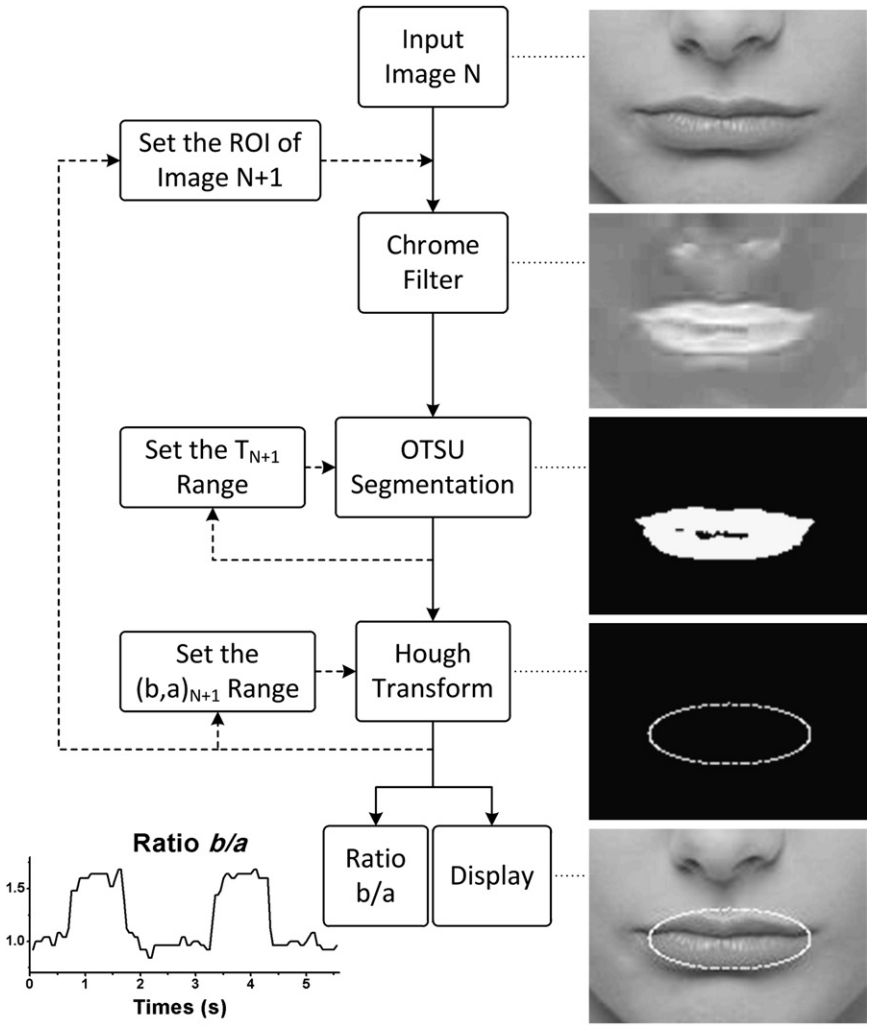


**FIGURE 2.** The procedure for extraction of lip outer contour features. Images in the right column show the intermediate results corresponding to the process in the left column.

is used as a measure of lip deformation. The lips become longer longitudinally and remain invariant or shortened latitudinally when voicing most phonemes, which results in an increase in the ratio *b*/*a*. Therefore, the user is regarded as speaking when the ratio of *b*/*a* exceeds a certain threshold. However, a simple threshold results in an intermittent voice when speaking continuously because of some close-mouth phonemes and transitional lip shapes. It is difficult to distinguish such lip shapes from silence by only measuring the ratio *b*/*a*. Hence, we applied an off-delay time to improve fluency. When the ratio *b*/*a* decreases below a certain threshold, the on switch is to be held for an additional time of *m* ms until no more phonation is detected during this period. Based on our experimental knowledge of practical voicing, the threshold was set to 135% of the baseline, and the off-delay was set to 300 ms in this study.

## Implementation of a video-based experimental system

The experimental system (video-EL) is built as shown in Figure 3. A camera (Lenovo, China) is used to capture lip image sequences with a resolution of $480 \times 320$ in 20 FPS. The camera is fixed onto a microphone headset via an adjustable metal support. This fixture keeps the camera in a fixed position relative to the lips. When a subject puts on the headset, the camera is about 10 cm away from the lips, and the microphone is at the side of the cheek, approximately 2–3 cm away from the mouth to avoid blocking the camera. With this headset, the user can move his head during speaking without affecting the tracking of lip deformation. The captured image sequences are transmitted into a computer (KaiTian A6000; Lenovo, Beijing, China), which is used as a data processor. A computer program was written with VC++ to process the images from the video camera. The program includes optional functions for saving data, adjusting parameters, and displaying video and audio signals as visual feedback. The on/off command is outputted via a parallel port to a driving circuit that triggers a wearable EL vibrator (Neck-Type; XinYu, Daqing, China).

## Subjects

Eight subjects, including one laryngectomee and seven subjects with normal voice, were invited to use the video-EL for method assessment. The laryngectomy subject was aged 74 years, had


**FIGURE 3.** Implementation of the video-EL system. The upper image shows the composition of video-EL, and the lower image shows a subject using video-EL.

a larynx removal 3 years ago, and used an EL for speech rehabilitation for more than 2 years. The ages of the seven subjects with normal voice (four men and three women) ranged from 24 to 28 years, and the average age was 25 years. Three of them were skilled EL speakers who could produce high-quality electrolaryngeal speech with a button-EL. All the subjects were native Mandarin Chinese speakers without any facial or oral cavity abnormality. Before experiments, subjects were first introduced to the working principle of the video-EL system and were then given approximately 5–10 minutes training time to use the video-EL.

## Simulation behaviors

A simulated voicing experiment was adopted to assess the current parameter setting (threshold = 135% of baseline; off-delay time = 300 ms) by comparing with the optimal parameter obtained by minimizing the total switching errors. The ratio *b*/*a* was used for simulating the output of the video-EL at different threshold levels and off-delay times. Errors, including those of type I and type II for mistaking silence for voicing and mistaking voicing for silence, respectively, were calculated by comparing the simulated output with the audio signals.

The seven subjects with normal voice participated in the test. The laryngectomy subject was absent for the lack of normal voice. First, the subject was asked to keep silent for approximately 1 minute to acquire a baseline. Then, the subject was asked to read a paragraph without feedback. The paragraph "The North Wind and the Sun" consisting of 160 words, 15 commas, and seven full stops was chosen because it contained almost all the Mandarin phonemes. Subjects read the paragraph with their normal speaking rate three times. The ratio *b*/*a* and the audio signal were collected simultaneously during speaking.

A variable threshold level and an alternative off-delay time were applied to simulate the output of the video-EL. Type I error duration was normalized by the total time of the silent period, and type II error duration was normalized by the total time of the voicing period. The total error was counted as the sum of type I and type II errors. In addition, the total number of occurrences of type II error was counted. The current parameter settings were compared with the threshold that resulted in the minimum total error.

## Reaction time test

A reaction time test was used to compare the onset and offset performance of video-EL with normal voice and button-EL. The reaction time was defined as the time difference between an action cue and voice initiation or termination.

The experimental setup is shown in Figure 4. A light-emitting diode (LED) was used as a visual cue of action. A high-potential stimulus generated by LabVIEW (National Instruments, Austin, TX) was used to light the LED via a data acquisition board (PCI6023e; National Instruments). A microphone was used to collect an audio signal. The audio signal and LED stimulus were recorded by a multiple-channel signal acquisition system (MP150; BIOPAC, Santa Barbara, CA) simultaneously.

All eight subjects participated in the test. Subjects were asked to pronounce five vowels (*a*, *i*, *e*, *o*, *u*) and four single syllables of consonant-vowel (C-V) structure (*ha*, *fa*, *ta*, *ma*) with normal voice, button-EL, and video-EL. The vowels were selected for their central role in speech production (especially in Mandarin Chinese, which is a monosyllabic language). The syllables were used to estimate the influence that different types of consonants had on voice initiation.

Within each trial, a sequence of visual cues was presented to guide the subject. First, the subject was required to stay in a resting state for approximately 10 seconds. Then, a pop-up message asked the subject to get ready for voicing. The "get ready" period was chosen as a random length of 1–2 seconds to reduce the

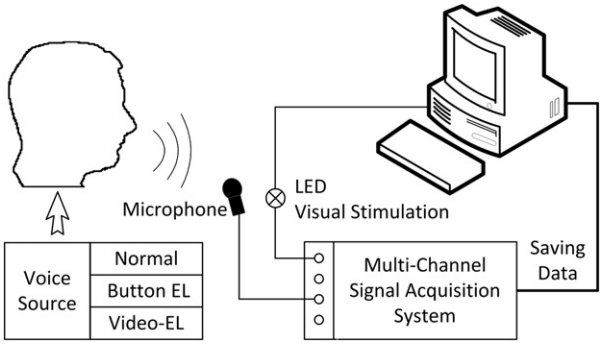


**FIGURE 4.** Setup for measuring reaction time. A visual cue is produced by an LED controlled by the computer via a data acquisition board. The audio signal and LED stimulus are recorded by a multichannel signal acquisition system simultaneously.

effect of the subject's anticipation and to minimize the reaction time.[21] This "get ready" period followed the action cue of the LED. The subject was asked to start voicing as soon as the action cue appeared. The voicing period was also a period of random length of 2–4 seconds. At the end of the voicing period, the LED was turned off as the cue to stop voicing. During each round, the subject produced voices in the order *a*, *i*, *e*, *o*, *u*, *ha*, *fa*, *ta*, *ma* with one voice source. The subject was allowed to take a break after each round. Each voice source was tested 10 times. The order of testing of the different voice sources was randomized for each subject.

The voice initiation time (VIT) was calculated as the time difference between the action cue and the start of speech, and the voice termination time (VTT) was calculated as the time difference between the stop cue and the end of speech. The mean VIT and VTT for each voice source were used for statistical analysis.

### Fluency test

Subjects were asked to read several paragraphs with video-EL. The speech was played to listeners, and the errors in speech were labeled and counted to assess the fluency of video-EL.

Four subjects, including the laryngectomee and three normal subjects who could produce high-quality electrolaryngeal speech, were invited to the fluency test so that listeners could score the speech accurately. Subjects were asked to read several paragraphs in their normal speaking rate. A window displaying the ratio *b*/*a* was shown on the monitor as visual feedback. The audio signal was picked up with a high-quality microphone and saved in MP3 file format at a rate of 44 100 samples per second.

Ten classical Chinese poems, including five five-character poems and five seven-character poems, were chosen because their rhyme and prosodic pattern induced a melodic cadence.[13] Each poem consisted of four lines; there were $(5 \times 4 + 7 \times 4) \times 5 = 240$ words and $4 \times (5 + 5) = 40$ stops in all. To avoid the influence of reading problems in classical Chinese words, all the poems were given to subjects in advance.

All the sounds were judged by the first author. The MP3 files were played to the listener, and the waveforms were shown to assist in evaluating the speech. Sounds that stopped within sentences or were produced in breaks were labeled as four types of errors: incorrect onset, devoicing midword, unfinished word, and error voicing in breaking. The first three error types appeared within sentences and were distinguished by the position of the stops. A stop that led to the front part of a word being devoiced or delayed was counted as an incorrect onset. A stop within a word was counted as a devoicing midword. A stop that led to the back part of a word being devoiced was counted as an unfinished word. Errors voicing only appeared in the breaks and included perceivable extended endings and voicing during breaks. Accordingly, two kinds of scores were calculated, one being the total errors in sentences, which were calculated as the percentage of the first three types of errors within total words, and the other being the total errors in stops, which were calculated as the percentage of error voicing within total stops.

The first author judged all the trials and repeated a subset (20% of the total), which was randomly picked up from all files. Another listener who was familiar with electrolaryngeal speech judged the same subset. The second listener's judgments were compared with that of the first listener to calculate the reliability of scoring. The judgment of fluency of the first author had a reliability score of 99%. The subset scored by both the first author and the second listener had a reliability score of 92%.

### Intelligibility test

The intelligibility of video-EL speech was measured by scoring a group of words and sentences via listener comprehension and was compared with button-EL to assess the influence of voice initiation/termination.

The four subjects who participated in the fluency test were also invited to the intelligibility test for their capability to produce high-quality electrolaryngeal speech. Each subject was asked to read reading materials with video-EL and button-EL. Reading material contained four categories: vowels, syllables with C-V structure, syllables with C-V-C structure, and sentences. The first category contained 10 Chinese vowels, including single and compound vowels. The second and third categories were chosen from the China National Standards of Acoustic-Speech articulation testing method (GB/T 15508-1995). A table of syllables was randomly picked from 10 equivalent tables for each subject. Each table contained five vowels, 46 C-V syllables, and 24 C-V-C syllables. Except for the vowels being included in the first category, syllables were classified into second and third categories according to their structure. All the syllables were read once without carrier sentences. The fourth category contained eight sets of sentences, which were chosen from Chinese newspapers. Each set contained 10 sentences without quotations, reduplication, parentheses, or proper names more than three words. The sentences had an average length of 13.5 words (with a standard deviation [SD] of 2.2 words). Each subject randomly selected a set within the categories for each voice source. Each sentence was read once. Repetitions were allowed only if a nonlinguistic error occurred, such as a cough, sneeze, or an outside sound disturbance. All the sounds were recorded in MP3 file format at a rate of 44 100 samples per second. There were 660 isolated syllables $((10 + 75) \times 8 - 20)$ and 80 sentences for listener evaluation.

Eight listeners, who were all native Chinese speakers and did not have experience communicating with EL users, were invited to assess intelligibility. The audio files were presented independently to listeners via binaural earphones at a comfortable volume. Listeners were allowed to listen to the sounds as many times as needed. Listeners were asked to write down the target syllables and sentences. A correct response was counted as long as the right syllable was recorded. The intelligibility was calculated as the percentage of correct responses. The average intelligibility was analyzed across four categories.

## RESULTS

### Simulation behaviors

When the off-delay time was set to 0 ms, the average optimal threshold of all subjects was 120.4% of the baseline, with an average minimum total error of 33.1%. When the minimum

total error was achieved, the average type I and type II errors were 21.8% and 11.3%, respectively, and the type II error appeared on average 67.8 times. When the threshold and off-delay time were set to 135% and 300 ms, respectively, the average total error was 33.5%, which was not remarkably larger than the minimum total error under the optimal threshold. The average type I error and type II errors were 25.4% and 8.1%, respectively. The average appearance of type II errors decreased to 23 times. Figure 5 shows an example of simulation errors seen with one subject. Applying the off-delay time reduced the number of nontriggering (type II) errors by removing nontriggering errors of less than 300 ms and increased false triggering (type I) errors by adding extended triggering at the end of each triggering. In addition, the results showed that the type II error increases more slowly and the total error curve is much flatter from 120%–140% when an off-delay time is applied.

## Reaction time test

The VIT and VTT did not show significant differences across different vowels for each subject when the same voice source was used (*post hoc* tests). Therefore, the average VIT and VTT for each voice source were calculated as the mean value for all vowels. The average reaction times are shown in Figure 6. For both initiation and termination, *post hoc* tests showed that the reaction times for normal voice and button-EL were not significantly different from each other, and the reaction time for video-EL was significantly larger than normal voice and button-EL. In the case of the laryngectomy subject, the VIT of video-EL was 546 ms, which was slightly larger than that of button-EL (493 ms), and the VTT for video-EL was 833 ms, which was much larger than that of button-EL (419 ms).

When producing syllables with an initial consonant, four types of consonants showed different influences on voice initiation. *Post hoc* tests showed that the VIT of normal subjects pronouncing *ha* and *ma* with normal voice was smaller than that of the vowel *a*, and the VIT of normal subjects pronouncing *fa* and *ta* was not significantly different from that of the vowel *a*. The VIT of normal subjects producing speech (including syllables and the vowel *a*) with button-EL was not significantly different from each other. When subjects used video-EL, the VIT of the syllable *ha* was not significantly different from the vowel *a*, and the VIT of the syllables *fa*, *ta*, and *ma* was larger than that of the vowel *a*. Using button-EL, the laryngectomy subject had a VIT that was 31 ms less and 99, 58, and 33 ms larger than the vowel *a* when pronouncing *ha*, *fa*, *ta*, and *ma*, respectively. Using video-EL, the laryngectomy subject had a VIT that was

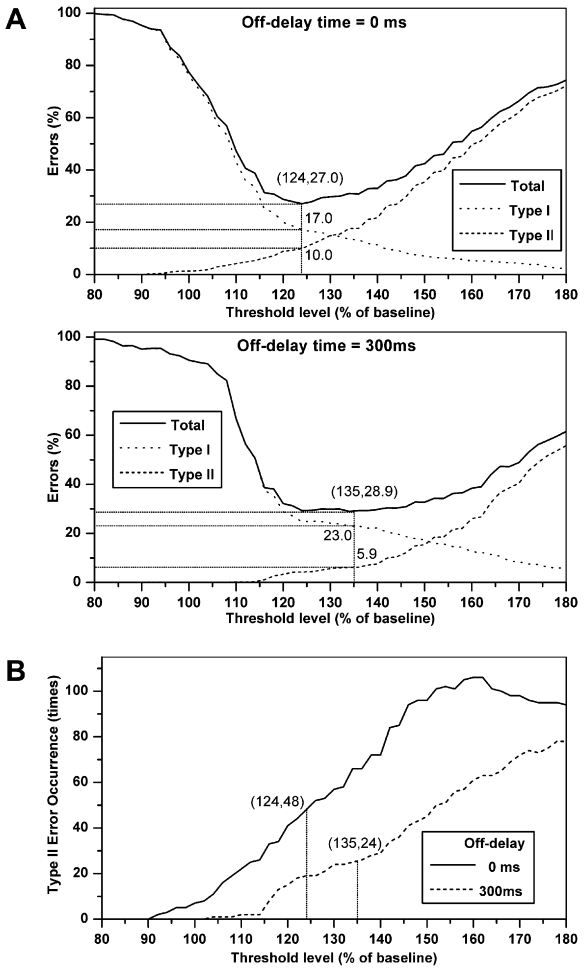


**FIGURE 5.** The simulation output errors from one subject with normal speech. A. Errors vary with the threshold level under 0 ms and 300 ms off-delay time. B. The number of type II error occurrences vary with the threshold level.

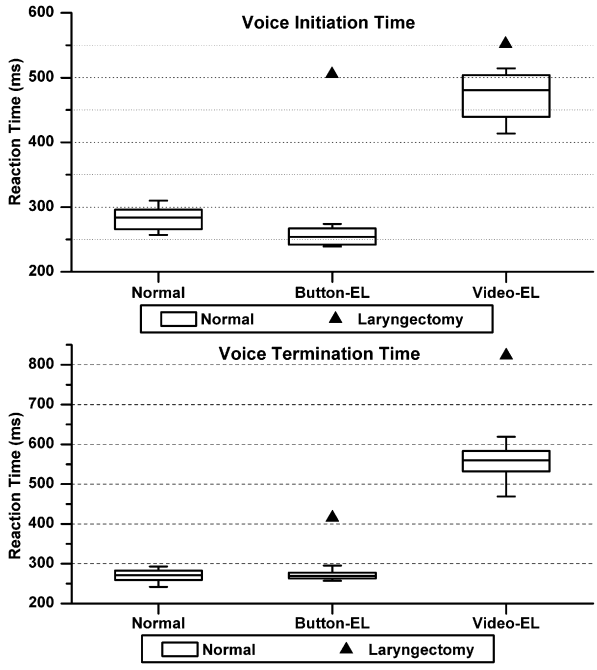


**FIGURE 6.** Voice initiation and termination times for subjects pronouncing vowels. Box plots show the median, interquartile, and min-max values of VIT and VTT for seven subjects with normal speech. Solid triangles show the reaction time of the laryngectomy subject using button-EL and video-EL. Data for normal voice are absent for the lack of larynx.

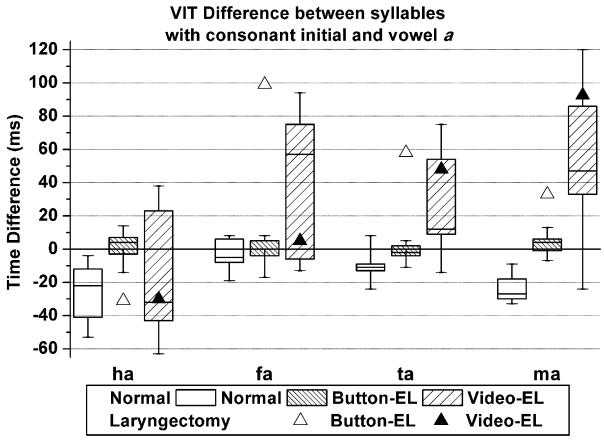


**FIGURE 7.** Time difference between initiations of syllables with an initial consonant and initiation of the vowel *a*. The three kinds of boxes show the VIT difference of normal subjects using normal voice, button-EL, and video-EL, respectively. The triangles show the time difference of the laryngectomy subject using button-EL and video-EL. Positive values indicate that the syllable starts slower than the vowel *a*, whereas negative values indicate that the syllable starts faster than the vowel *a*.

30 ms less and 5, 48, and 92 ms larger than the vowel *a* when pronouncing *ha*, *fa*, *ta*, and *ma*, respectively (Figure 7).

### Fluency test

The mean error in sentences produced by normal subjects was 4.2%, of which 84.3% were categorized as incorrect onset and 15.7% were categorized as devoicing midword; no unfinished words were found. The mean error in stops produced by normal subjects was 65%. In the case of the laryngectomy subject, the mean errors in sentences and in stops were 3.8% and 72.5%, respectively, and the number of instances of incorrect onset and devoicing midword within sentences were 77.8% and 22.2%, respectively. There was no significant difference between the fluency of the laryngectomy subject and subjects with normal speech (*t* tests; $P > 0.05$).

### Intelligibility test

The intelligibility of speech produced with button-EL and video-EL is shown in Figure 8. When producing single syllables with button-EL, the correct perception rate of vowels, C-V, and C-V-C syllables were 41.3%, 32.6%, and 14.6%, respectively. When producing syllables with video-EL, the correct perception rates were 29.9%, 4.8%, and 12.7%, respectively. Vowels and C-V syllables produced with video-EL were significantly less intelligible than those produced with button-EL (*t* test; $P < 0.001$). The intelligibility of video-EL was not significantly different from that of button-EL when pronouncing C-V-C syllables (*t* test; $P = 0.353$). The percentage of correct perception of initial consonants (in categories 2 and 3) were 33.1% and 32.3% when using button-EL and video-EL, respectively, and there was no significant difference between the two (*t* test; $P = 0.572$).

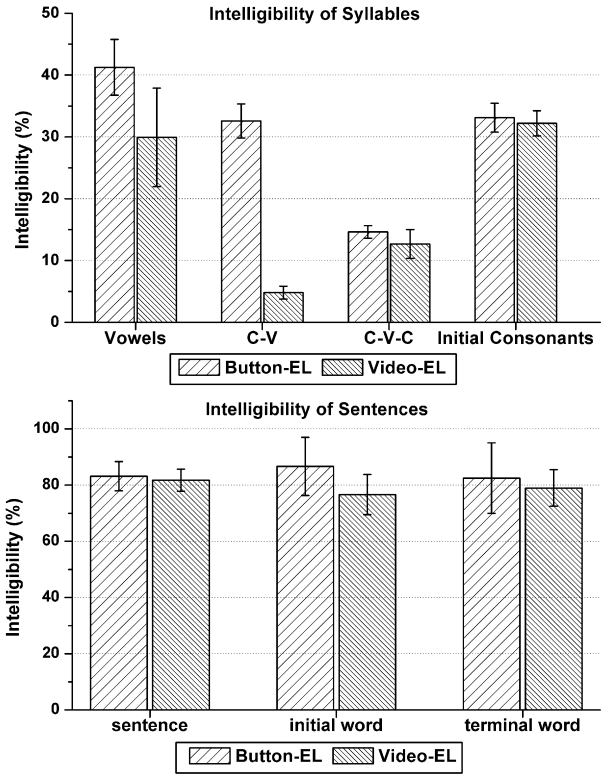


**FIGURE 8.** Intelligibility score of speech produced with button-EL and video-EL. The histogram shows the average intelligibility for all listeners when listening to pronounced syllables or sentences. The error bar shows the standard deviation.

When producing sentences with button-EL and video-EL, the intelligibilities were 83.2% and 82.5%, respectively. The intelligibility of initial words of sentences was 86.6% and 78.6% for button-EL and video-EL, whereas the intelligibility of terminal words was 82.5% and 78.9%, when speaking with button-EL and video-EL, respectively. One-way analysis of variance tests showed that there were no significant differences in the intelligibility of sentences, initial words, and terminal words for each EL device. There was also no significant difference between speech produced with button-EL and video-EL in each group (*t* test; $P > 0.05$).

## DISCUSSION

### Parameter settings

The current parameter settings (135% threshold and a delay of 300 ms) do not show significant difference with the optimal threshold setting by comparing the total error. Although false triggering errors become larger in either case, the nontriggering error decreases remarkably when the current parameter setting is used. The current parameter settings result in nearly the same total error as the optimal threshold settings. Interruptions within words or sentences are considered to have more influence on communication than unintended triggering of speech. Therefore, type II errors are not weighted the same as type I errors for perception of voice. Adjusting the relative weight between the two types of errors results in different optimal thresholds.

Regardless of the off-delay, increasing the relative cost of type I to type II errors produces larger optimal thresholds, and decreasing the relative cost produces smaller optimal thresholds. When the optimal threshold is 135% of the baseline (with an off-delay of 300 ms), the average relative cost of type I to type II errors for all subjects is 1.33. This result is much larger than the relative cost found previously by Goldstein et al.[11] This seems to imply an opposite inference; users prefer to decrease unintended triggering rather than nontriggering errors. The contradiction can be explained as follows. Firstly, the off-delay time makes the error curve much flatter in the interval of 120%–140%. Nontriggering errors that occur during intended voicing do not increase much in this interval. In addition, original false triggering in silence and extended triggering brought on by an off-delay time does not have the same weight. However, type I errors, which are the sum of the two types of errors mentioned above, do not change much in the same interval. Therefore, as the threshold increases, the false triggering in silence decreases and the extended triggering brought on by an off-delay time increases. Thus, the relative cost actually indicates that users are more concerned about reducing the original false triggering than extended triggering brought on by an off-delay time.

In this experiment, subjects produced voice without any feedback, which would certainly affect users' behavior. Thus, the simulation did not represent the error level of practical voicing. It was observed that subjects often kept their mouths open during short breaks within sentences, which induced false triggering during silence. Therefore, it is believed that subjects could produce fewer errors if feedback were provided during practical voicing. Practical voicing will be discussed in the subsequent sections.

## Switching performance

The results from the reaction time test and the fluency test evaluated the switching performance of video-EL-produced single words and long sentences, respectively. The reaction time of normal subjects using normal voice (VIT = 282 ± 21.6 ms; VTT = 269 ± 20.0 ms) and button-EL (VIT = 255 ± 15.3 ms; VTT = 272 ± 14.9 ms) was very close to the results of previous reports.[11,21,22] Therefore, the setup and procedure that was used to collect the reaction time data were deemed valid. Subjects using video-EL had larger VIT than those using other methods including EMG-EL. The reason may be because of (1) the time required for video processing and (2) the hysteresis caused by the threshold, for that the vibrator is not triggered until the lips deform over a certain degree. Thus, the onset delay for video-EL was larger than normal voicing, which only requires an adjustment of the larynx muscles and a buildup of subglottal pressure; button-EL, which only requires the user to press a button; and EMG-EL, which is triggered by EMG activity that precedes the voice by 70–120 ms.[11,23] Subjects using video-EL had the largest VTT as well. This is mainly caused by the 300 ms off-delay, which results in a fixed hysteresis.

Initial consonants have different influences on voice initiation because of the different triggering manners of voice sources. Normal voice start-up is related to air pressure requirements. It takes less time to pronounce the unvoiced glottal fricative *h* or nasal *m* than vowels that require establishment of an underglottal pressure or the unvoiced labiodental fricative *f* and explodent *t* that require certain intraoral pressures. Button-EL is controlled by hand, which is related to muscular activity; thus, the VIT for button-EL is not significantly different across syllables. For video-EL, the initiation is determined by lip activity only. The lip deformation hardly exceeded the threshold when certain close-mouth phonemes were pronounced, such that video-EL was not triggered until mouth opening. Therefore, syllables with initial close-mouth phonemes, such as *f*, *t*, and *m*, were triggered slower than syllables with initial open-mouth phonemes, such as *h* and *a*, and this delay might be related to the voice onset time because of the phonation processing. Moreover, the SD of VIT produced with video-EL was much bigger than that with normal voice and button-EL. This might be caused by the individual differences among subjects, such as lip shape, articulatory movements, or articulatory habits during voicing, which do not seriously affect the initiation of normal voice or button-EL speech but has a direct impact on the triggering of video-EL based on lip motion. This large SD also implies that lip motion does not have such a strict correspondence with phonemes as other factors. Therefore, there were always some subjects who produced syllables with initial close-mouth phonemes earlier than single vowels, whereas others produced syllables with initial close-mouth phonemes later than single vowels. This implies that it may be possible to reach faster initiation without affecting phonation by targeting training on lip control.

In the case of the laryngectomy subject, both VIT and VTT were much larger than that of normal subjects when voicing vowels with button-EL or video-EL. This larger reaction time might be related to several factors, including individual reaction rate, habitual phonation processing with EL, or reaction to the test procedure. However, it is hard to verify which factor was the most influential based on the results of only one subject. Despite this, the reaction time difference between video-EL and button-EL for the laryngectomy subject still revealed the different performances of the two triggering methods. These results indicate that the laryngectomy subject pressed the button with his finger using button-EL earlier than making any detectable lip deformation, as was the case with normal subjects. When voicing syllables with video-EL, the VIT difference of each syllable was not significantly different from that of normal subjects (*t* tests; $P > 0.05$). This seems to imply that the removal of the larynx has no influence on the effect of initial consonants on voice initiation.

When speaking long sentences, all the subjects produced fluent speech with video-EL, and this was accompanied by a small error rate. The error was much less than the simulation output errors mentioned above. This might be caused by the effect of audio and visual feedbacks. Lip motion of most voices triggered the EL vibrator, and the off-delay time made up the interval between two triggers. The most frequent error in sentences was incorrect onset, which made up approximately four in five of the total errors; the remaining was devoicing midword errors, which made up approximately one in five of the total errors. Unfinished word errors were not found among any of the subjects.

Almost all the errors in sentences occurred when the word contained the single vowel *i* followed by a word with close-mouth initiation. It was found that all the vowels triggered the video-EL on time except for the single vowel *i*. This might be because of the fact that the lips exhibit less deformation when pronouncing the vowel *i*. Users have to adjust their lips according to feedback (the break within a sentence) to reproduce the voice. The fact that there were no unfinished words was mainly because of the off-delay time. An off-delay time of 300 ms is sufficient to finish a Chinese word, for that the average speaking time per word is 224 ms during the normal speaking rate for Mandarin.[24]

The off-delay ensured that a single word finished correctly and a sentence was uninterrupted in most situations but caused a perceptible extended triggering at the end of a sentence. However, users prefer fluent voice more than noiseless breaks. Moreover, as mentioned above, users are more concerned about reducing original false triggering errors than extended triggering brought on by the off-delay time. It has to be acknowledged that the off-delay is not helpful for words at the beginning of a sentence and induces an extended ending. The influence of this on speech intelligibility is discussed later.

Speaking rate is also an important factor that influenced the performance of video-EL. Speaking rate affects lip motions, including the speed and amplitude of lip deformations,[19,20] which affects the triggering of voicing. In addition, the speaking rate influences the effect of the off-delay time. With a certain off-delay time, fast speaking rates lead to long unintended voicing at the end of a sentence, whereas slow speaking rates do not allow for effective removal of errors in sentences. In this experiment, subjects were asked to voice with their normal speaking rate. Although video-EL with the current setting already provides a fluent voice source for subjects voicing with their normal speaking rate, better performance may be achieved by training users to speak with an optimal speaking rate or by finding the optimal off-delay setting according to individual speaking rate. Further study will be focused on such training.

### Speech intelligibility

By comparison with button-EL, the speech intelligibility test showed the influence that video-EL had on rehabilitated speech. Because of the working principle of video-EL, voice produced with video-EL had two shortcomings: slow voice initiation and fixed extended ending. When isolated syllables were pronounced, excluding C-V-C syllables, the intelligibility of video-EL speech was remarkably less than that of button-EL speech. When pronouncing C-V-C syllables, the intelligibility of video-EL speech was not significantly different than that of button-EL speech. It was also found that almost all the single vowels were misinterpreted as nasal rhymes when using video-EL. For example, *i* was often perceived as *in* or *ing* and *e* was perceived as *en* or *eng*. This is obviously because of the extended ending that produces an additional nasal terminal on the closing of users' mouths. This may be a major reason for the reduction of intelligibility of vowels and C-V syllables produced with video-EL. In Chinese, C-V-C syllables can only be ended with nasal rhymes $-n$ or $-ng$. Thus, the extended ending does not seriously affect the perception of these nasal rhymes. The intelligibility of initial consonants was not significantly different between button-EL and video-EL either. Button-EL provides a stable voiced source for pronouncing initial consonants, whereas video-EL provides a variable voiced source relating to phonemes, vocalizing manners, or individual characteristics. For electrolaryngeal speech, it has been reported that word initial consonants (particularly voiceless stops) are the least intelligible phonemes.[25] The voiced source may induce confusions of a voiceless consonant with its cognate. However, for video-EL speech, it is hard to validate the specific influence of each voicing condition because of large differences between individuals. Despite the uncertainty about how video-EL influences intelligibility, our results imply that the role of voice initiation of video-EL is less important than the phoneme position or the voiced source.

The intelligibility between video-EL and button-EL speech does not show significant differences when sentences are produced. This is mainly because of contextual cues that are helpful for listeners understanding EL speech.[26] Slow voice initiation and fixed extended ending only appear once in one sentence if the sentence is fluently uttered. In this study, we have shown that video-EL can provide a fluent voice source with few breaks within a sentence. When taken together, the results from the fluency and intelligibility tests may imply that contextual information may compensate for the loss of speech quality caused by the flaws of voice source produced with video-EL. In other words, video-EL could provide a sufficiently fluent voice source so that the intelligibility of speech is not seriously affected by switching errors.

## CONCLUSION

In this article, we validated a method for the automatic on/off control of an EL. Lip deformation was used as the controlling source. An experimental EL system based on real-time video processing (video-EL) was implemented and its performance in producing Mandarin Chinese was assessed. Video-EL could not generate voice initiation and termination as soon as a button-EL and affected the perception of isolated word. However, when producing sentences, the video-EL system provided a fluent voice source and electrolaryngeal speech that was as intelligible as that produced with a button-EL.

This method is proved as an effective technique for the automatic on/off control of an EL. However, the current requirement of a computer in the experimental system restricts the application of video-EL in daily life. A portable video-EL device is being developed to meet daily communication needs.

### Acknowledgments

The authors deeply appreciate Ye Wenyuan and Rui Baozhen for their cooperation in the experiments. This work was supported in part by the National Natural Science Foundation of China under grant 30770544, grant 10874137, and Key Project of National Basic Research Development Program of China (973 Program) under grant 2010CB732603.

## REFERENCES

1. Lauder E. The laryngectomee and the artificial larynx–a second look. *J Speech Hear Disord*. 1970;35:62–65.

2. Rothman H. Acoustic analysis of artificial electronic larynx speech. In: Seikey A, ed. *Electroacoustics Analysis and Enhancement of Alaryngeal Speech*. Springfield, IL: Charles Thomas; 1982:95–118.
3. Liu H, Ng ML. Electrolarynx in voice rehabilitation. *Auris Nasus Larynx*. 2007;34:327–332.
4. Meltzner GS, Hillman RE, Heaton JT, Houston KM, Kobler JB, Qi Y. Electrolaryngeal speech: The state of the art and future directions for development. In: Doyle PC, Keith RL, eds. *Contemporary Considerations in the Treatment and Rehabilitation of Head and Neck Cancer: Voice, Speech, and Swallowing*. Austin, TX: Pro-Ed; 2005:571–590.
5. Zwitman DH, Knorr SG. The design of a wireless-controlled intra-oral electrolarynx. *J Bioeng*. 1977;1:165–171.
6. Zwitman DH, Knorr SG, Sonderman JC. Development and testing of an intraoral electrolarynx for laryngectomy patients. *J Speech Hear Disord*. 1978;43:263–269.
7. 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.
8. Hashiba M, Sugai Y, Lzumi L, Ino S, and Ifukube T. Development of a wearable electrolarynx for laryngectomees and its evaluation. In: *Proceeding of the 29th Annual International Conference of the IEEE EMBS*; Lyon, 2007: 5267–5270.
9. Heaton JT, Kobler JB, Goldstein EA, Mcmahon TA, Barry DT, Hillman RE. Recurrent laryngeal nerve transposition in guinea pigs. *Ann Otol Rhinol Laryngol*. 2000;109:972–980.
10. Heaton JT, et al. Surface electromyographic activity in total laryngectomy patients following laryngeal nerve transfer to neck strap muscles. *Ann Otol Rhinol Laryngol*. 2004;113:754–764.
11. 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 Trans Biomed Eng*. 2004; 51:325–332.
12. Goldstein EA, Heaton JT, Stepp CE, Hillman RE. Training Effects on Speech Production Using a Hands-Free Electromyographically Controlled Electrolarynx. *J Speech Hear Res*. 2007;50:335–351.
13. Kubert HL, et al. Electromyographic control of a hands-free electrolarynx using neck strap muscles. *J Commun Disord*. 2009;42:211–225.
14. Stepp CE, Heaton JT, Rolland RG, Hillman RE. Neck and face surface electromyography for prosthetic voice control after total laryngectomy. *IEEE Trans Neural Syst Rehabil Eng*. 2009;17:146–155.
15. Hennecke ME, Prasad KV, Stork DG. Using deformable templates to infer visual speech dynamics. *28th Asilomar Conf Signals Syst Comput*. 1994;1: 578–582.
16. Zou Y, Wang B. Lip-tracking algorithm based on elliptical deformable templates. *Chin J Scientific Instrument*. 2007;28:514–518.
17. Yao H, Lu Y, Gao W. Lip-Movement Features Extraction and Recognition Based on Chroma Analysis. *ACTA Electronica Sinca*. 2002;30:168–172.
18. Otsu N. A threshold selection method from gray-level histograms. *IEEE Trans Syst Man Cybern*. 1979;9:62–66.
19. Goffman L, Smith A. Development and phonetic differentiation of speech movement patterns. *J Exp Psychol Hum Perception Perform*. 1999;25: 649–660.
20. McClean MD, Tasko SM. Association of orofacial muscle activity and movement during changes in speech rate and intensity. *J Speech Lang Hear Res*. 2003;46:1387–1400.
21. Watson B. Foreperiod duration, range, and ordering effects on acoustic LRT in normal speakers. *J Voice*. 1994;8:248–254.
22. Cullinan WL, Springer MT. Voice initiation and termination times in stuttering and nonstuttering children. *J Speech Hear Res*. 1980;23: 344–360.
23. Hilel AD. The study of laryngeal muscle activity in normal human subjects and in patients with laryngeal dystonia using multiple fine-wire electromyography. *Laryngoscope*. 2001;111:1–47.
24. Lee PSK, Chan AHS. Chinese speaking times. *Int J Ind Ergon*. 2003;31: 313–321.
25. Weiss MS, Basili AG. Electrolaryngeal speech produced by laryngectomized subjects: perceptual characteristics. *J Speech Hear Res*. 1985;28: 294–300.
26. Liu H, Wan M, Wang S. Features of listeners affecting the perceptions of Mandarin electrolaryngeal speech. *Folia Phoniatr Logop*. 2005;57:9–19.