Development and Evaluation of On/Off Control for Electrolaryngeal Speech Via Artificial Neural Network Based on Visual Information of Lips

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

Summary: Objective. To realize an accurate and automatic on/off control of electrolarynx (EL), an artificial neural network (ANN) was introduced for switch identification based on visual information of lips and implemented by an experimental system (ANN-EL). The objective was to confirm the feasibility of the ANN method and evaluate the performance of ANN-EL in Mandarin speech.

Study Design and Methods. Totally five volunteers (one laryngectomee and four normal speakers) participated in the whole process of experiments. First, trained ANN was tested to assess switch identification performance of ANN method. Then, voice initiation/termination time, speech fluency, and word intelligibility were measured and compared with button-EL and video-EL to evaluate on/off control performance of ANN-EL.

Results. The test showed that ANN method performed accurate switch identification (>99%). ANN-EL was as fast as normal voice and button-EL in onset control, but a little slower in offset control. ANN-EL could provide a fluent voice source with rare breaks (<1%) for a continuous speech. The results also indicated that on/off control performance of ANN-EL had a significant impact on perception, lowering the word intelligibility compared with button-EL. However, the words produced by ANN-EL were more intelligible than video-EL by approximately 20%.

Conclusions. The ANN method was proved feasible and effective for switch identification based on visual information of lips. The ANN-EL could provide an accurate on/off control for fluent Mandarin speech.

Key Words: Artificial neural network–Electrolarynx–On/off control–Visual information.

INTRODUCTION

Speech is the most important and efficient way of communication. Owing to laryngeal cancer or trauma, people are prone to remove their entire larynx and, therefore, lose their physiological structure for normal speech. However, taking advantage of the remaining vocal tract and principle of speech production, electrolarynx (EL) speech is an effective way for voice rehabilitation and alaryngeal communication.

The EL is a handheld and battery-powered device, which transmits mechanical vibration into laryngopharynx through the neck or into posterior oral cavity with a tube or denture.[1] Owing to easy learning and no additional surgery required, EL speech has been widely accepted for daily communication by more than one-half of the laryngectomees.[1–3] However, the conventional EL is not convenient for users. The occupation of one hand to hold EL and control an on/off button during speech is ranked in the top five deficits of EL communication.[4] Therefore, many methods have been reported on switch control for hands-free EL.

There are three representative methods used for on/off control of EL without hands: tongue, electromyography (EMG), and visual information (video) control. Knorr, Zwitman, and colleagues[5,6] designed a wireless intraoral EL, which was switched on/off by the tongue. Although hands were not required for holding the device and pushing the button, users still had to control on/off actively. Goldstein et al[7] found that it was feasible to control initiation and termination of EL voice automatically by EMG signals from neck strap muscles. EMG-EL realized the hands-free control and was easy to master after proper training.[8] However, the biggest disadvantage of EMG control was an additional surgery to preserve the omohyoid strap muscles.[9] This would result in more pain and expenditure to the patients. To address this problem, Stepp et al[10] used neck and face surface EMG (sEMG) to control onset and offset of EL, and found that individuals were able to use sEMG from multiple recording locations to produce running speech perceived as natural as that produced with a typical handheld EL.

Visual information, especially the shape information of lips, has been extensively used in speech recognition,[11,12] speaker identification,[13,14] and perceptual evaluation[15,16] because of its close relationship with speech production. Recently, a noncontact method based on lip deformation was proposed for automatic on/off control of an EL (video-EL) by Wan et al.[17] The shape of lip outer contour was extracted through real-time video processing and presented by an ellipse with two parameters, namely the semimajor (a) and semiminor axes (b). Finally, the ratio of b to a (b/a) was used to determine switch on/off through a single threshold judgment. Wan et al[17] reported that video-EL was effective in the automatic on/off control and could produce fluent Mandarin speech as intelligible as button-EL. However, owing to the single parameter (b/a) and single threshold used in b/a method, video-EL could not generate voice initiation and termination as fast as button-EL, which affected the perception of isolated word. First, only one parameter (b/a) is limited to represent and differentiate all the lip shapes of phonation from silence. Second, the

Accepted for publication October 22, 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 Mingxi Wan or Supin Wang, 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. E-mail: mxwan@mail.xjtu.edu.cn or spwang@mail.xjtu.edu.cn

Journal of Voice, Vol. 27, No. 2, pp. 259.e7-259.e16
0892-1997/$36.00

http://dx.doi.org/10.1016/j.jvoice.2012.10.011

parameters (b/a) of phonation and silence are not linearity separable, so linear classification with a fixed single threshold could not absolutely distinguish one from the other, such as some closed-mouth phonemes mentioned by Wan et al.[17]

To realize an accurate voice initiation and termination, an artificial neural network (ANN) was introduced for switch identification and on/off control based on visual information of lips. ANN is a mathematical model widely applied in statistical pattern recognition.[18] The nonlinear nature of ANN will satisfy the mapping between lip features and voice on/off. Besides, the ANN has a strong robustness against noises. In this article, we implemented ANN method in a new video-controlled EL system (ANN-EL), which captured visual information of lips and controlled on/off of a wearable EL in real time. Furthermore, the performance of ANN method and ANN-EL were evaluated and compared with normal voice, button-EL, and video-EL.

METHODS

A schematic diagram of the experimental EL (ANN-EL) system is shown in Figure 1. The video signal of lips was captured and processed in real time to control on/off of EL. The procedure contained two main steps, which were lip-parameter extraction and on/off control.

Extraction of visual information of lips

There are two approaches widely used for extracting visual features, namely image-and model-based approaches. Considering the computational complexity and real-time implementation, model-based method was used to represent the lips by geometric parameters. The processes of parameters extraction were as follows: First, each frame of facial video image was preprocessed to decrease background noise and illumination. Second, the color image was filtered by a chromatic operator of lips and transformed to gray-scale image,[19] from which the lips was segmented by a threshold of gray-level histograms.[20] Finally, the lip outer contour was matched with an ellipse model and the shape parameters were extracted, namely semimajor (a) and semiminor axes (b).

On/off control with ANN

A two-layer feed-forward network was used in this article. It has been proved that with a sufficient number of hidden neurons, a multilayer perceptron neural network is capable of approximating an arbitrarily complex mapping within a finite support.[21] In the input layer, four inputs were normalized, namely semimajor axis (a/a_0), normalized semiminor axis (b/b_0), ratio of b to a (b/a), and normalized area of the ellipse (ab/a_0b_0). The parameters a_0 and b_0 represented the lip parameters of silence assigned during system initialization. Normalized parameters had advantages of distance and rotational invariances, so the influence of head movement could be reduced. For each neuron, the net function and the activation function were a weighted linear combination and a hyperbolic tangent activation function, respectively, which provided a nonlinear mapping between its input and output. The number of neurons was set as an empirical value of 20 in hidden layer. The network was trained using the scaled conjugate gradient back-propagation algorithm. The switch control depended on the two outputs, namely phonation and silence. The output of silence determined switch-off, and the other determined switch-on. Because the ANN algorithm was easy and fast, the real-time implementation was satisfied in our system.

The ANN-EL system

The ANN-EL system included three parts as shown in Figure 2. The first part was a microphone headset (Danyin DT-2699,

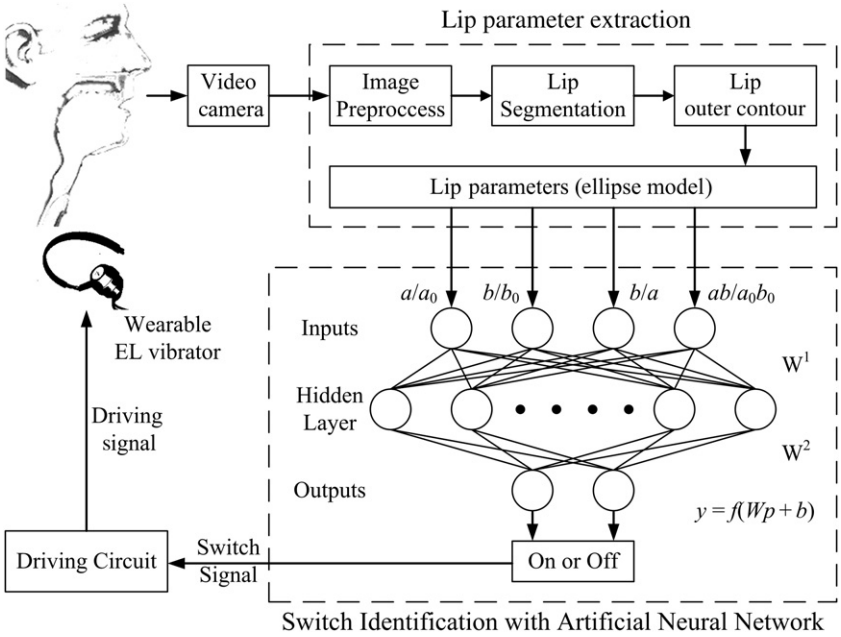

FIGURE 1. Schematic diagram of the ANN-EL system. The dash line boxes represent the extraction process of lip parameter and switch signal with artificial neural network.

FIGURE 2. The experimental electrolarynx (ANN-EL) system, which is used by an individual for hands-free EL speech.

Guangdong, China), which was modified to keep a camera (Lenovo, Beijing, China) in a fixed position to the lips via an adjustable metal support. This device can avoid the influence of head movement on video capture during speaking. When the headset was worn by an individual as shown in Figure 2, the camera was set about 10 cm in front of the lips, whereas the microphone was 5 cm away from the lips. The camera was used to capture the video of lips in 20 frames per second. The second part was a computer (KaiTian A6000, Lenovo), which was connected with camera via USB port. A program was developed to implement the before-mentioned methods and output switch signals. The last part was a driving circuit and a wearable EL vibrator (neck-type; XinYu, Daqing, China). The driving circuit received the switch signals via parallel port from the computer and then generated the driving signals to vibrate the EL.

EXPERIMENTS

Participants

One male laryngectomee and four normal speakers (two men and two women) participated in the following experiments. The laryngectomee, aged 74 years, underwent a total laryngectomy because of laryngeal cancer and had 2 years' experience of using button-EL. The normal subjects, aged from 22 to 26 years (averaging 24.25 years), had no reported history of speech language problems and were moderately proficient at using button-EL. All subjects were native Mandarin Chinese speakers.

Ten young adult listeners (five men and five women) aged from 22 to 28 years (averaging 25.3 years) participated in the perception task. All listeners had no reports of any hearing and language disorder and no previous exposure to alaryngeal speech.

ANN test

To train the ANN, sets of lip parameters (a and b) were collected as samples from normal subjects speaking with normal voice and laryngectomee with button-EL. All subjects were asked to pronounce sustained vowels (/ɑ/, /i/, /e/, /ɔ/, and /u/) after few seconds of silence. Meanwhile, the lip parameters were extracted and recorded synchronously with speech. In each recording, the mean values of parameters during the silence were set as a_0 and b_0, which were used to transform the lip parameters into the pairs of inputs and outputs for ANN. The data were selected as two types of samples (phonation and silence) according to speech signal, manually ruling out the situation as smile and yawn, and so on. Then, all selected data were randomly mixed and equally divided into two groups, one of which was used for ANN training and the other for ANN testing.

During the testing, two types of errors were calculated to evaluate the switch identification performance of ANN. Type I error was defined as the percentage of misidentified phonation samples, whereas type II error was the percentage of misidentified silence samples. Moreover, the errors were also calculated from b/a method using the empirical threshold (135% of b_0/a_0).[17] Then, the results of the two methods were compared in both normal subjects and laryngectomee.

Reaction time test

The reaction time experiment was adopted to evaluate on/off control performance of ANN-EL.[7,17] The experimental procedure was as follows: at the beginning, each subject sat in front of a display screen and had 10 seconds to relax. Then a "get ready" cue was displayed to indicate the subject to be ready for voicing. After a random 1–2 seconds, the subject was instructed to start voicing as soon as a "voicing" cue displayed on the screen. The voicing period lasted randomly for 2–4 seconds followed by a "stop" cue, which instructed the subject to stop voicing as soon as possible.

All the recordings were carried out in a soundproof room. The display (Lenovo) was placed 1 m away from the subject and controlled by *LabView* software (National Instrument). Speech and cue signals were collected synchronously using a data acquisition system (BioPac MP150) with a dynamic microphone (Salar M9, China) mounted 10 cm away from the mouth.

Two sets of materials were selected for reaction time test. One set contained five single vowels (/ɑ/, /e/, /i/, /ɔ/, and /u/), which were used for comparison with button-EL, video-EL, and EMG-EL. The other set contained 75 syllables (15 consonants $\times$ 5 vowels) with consonant-vowel (CV) structure,

which were used to estimate the influence of word-initial consonants on voice initiation time (VIT).

All subjects were asked to read the materials with four voice sources (ANN-EL, button-EL, video-EL, and normal voice). Each voice source was tested for 10 times by each subject. Finally, the VIT and voice termination time (VTT), standing for the time interval between the cue and voice initiation and termination, were measured and compared with different voice sources.

Fluency test

A passage entitled "Beifeng he Taiyang" (Boreas and Sun) was chosen for fluency test.[22] This passage contained all the Mandarin vowels and consonants, with 163 words and 18 pauses. All subjects were instructed to read this passage at a normal speed with ANN-EL and video-EL, respectively. Speech signals were collected with a dynamic microphone (Salar M9) mounted 10 cm away from the mouth.

Two cases were taken into account for fluency evaluation of EL speech. First case was unwanted break within a continuous phonation. Second case was unwanted voicing in the pause. Therefore, two corresponding errors were measured. One was break error in first case, which was measured as the percentage of words with undesired breaks. The other one was voicing error in second case, which was measured as the percentage of pauses without stopping voicing. The errors were judged subjectively according to the recording waveforms by all listeners, yielding a high interjudge reliability (97%).

Intelligibility test

A list of words was chosen from the National Standards of Peoples Republic of China: Acoustic-Speech articulation testing method (GB/T 15508-1995). The list contained 46 syllables with CV structure and 24 syllables with CVC structure. All subjects were asked to read the materials with ANN-EL, button-EL, and video-EL, respectively. EL Speech was recorded with a dynamic microphone (Salar M9) mounted 10 cm away from the mouth.

There were totally 15 sets (5 subjects × 3 voice sources) of 70 words recorded, which were played to listeners at a comfortable volume in the sound field of a quiet room. To avoid learning and experience effects, the order of words was set randomly. The listeners were instructed to transcribe the syllables using broad phonetic transcription. The intelligibility score was calculated as the mean percentage of correct responses to words for all listeners.

RESULTS

ANN test

The two types of errors for ANN and *b/a* method are shown in Figure 3. For normal subjects, the type I and type II errors of ANN method were $0.27 \pm 0.05\%$ and $0.47 \pm 0.07\%$, respectively. For laryngectomee, the type I and type II errors of ANN method were almost 0% and 0.3%, respectively. In contrast, the errors of *b/a* method were significantly larger than that of ANN method in all subjects (t test, $P < 0.05$, $N = 5$). The type I and type II errors of *b/a* method were $23.06 \pm 5.04\%$ and $6.01 \pm 4.27\%$ in normal subjects, whereas 19.71% and 7.2% in laryngectomee, respectively.

To reveal the differences between the two methods and different subjects, the confusion matrices[23] of one male normal subject and laryngectomee are listed in Tables 1 and 2, respectively. Each row represents the test samples in an actual class, whereas each column represents the samples in an identified class. For ANN method, the samples of phonation and silence were hardly misidentified (<1%), and there were no significant differences between individuals (analysis of variance [ANOVA], $P > 0.05$). For *b/a* method, 90.8% of /i/ samples were misidentified as silence in the normal subject, which was larger than 49.8% in laryngectomee. Moreover, 7.2% of silence samples misidentified as phonation in laryngectomee were much

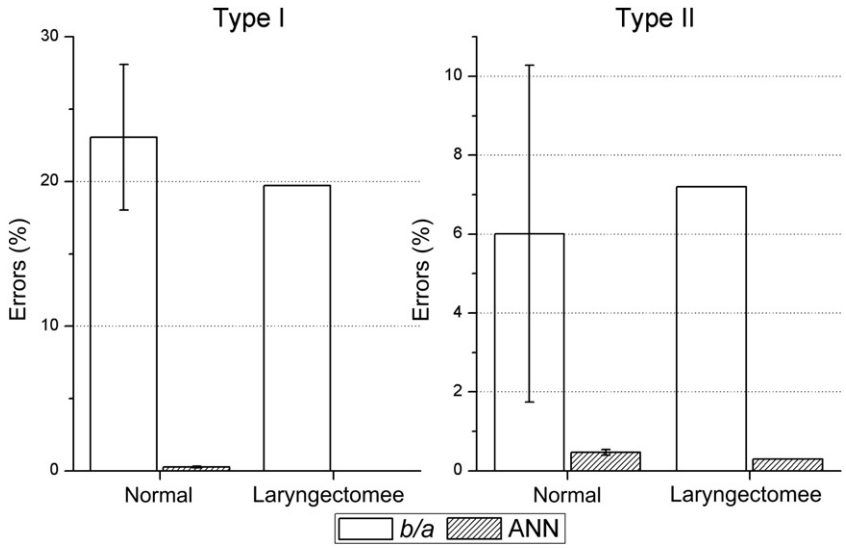

FIGURE 3. Two type errors of switch identification for ANN and *b/a* method in normal subjects and laryngectomee. The bars show the mean errors with standard deviation for four normal subjects and errors for a laryngectomee.

TABLE 1.
Confusion Matrix of Switch Identification for ANN and *b/a* Method in One Male Normal Subject

	Phonation		Silence	
Testing Samples	*b/a*, N (%)	ANN, N (%)	*b/a*, N (%)	ANN, N (%)
Phonation samples				
/ɑ/	5010 (100)	5010 (100)	0	0
/i/	483 (9.2)	5253 (100)	4770 (90.8)	0
/e/	5118 (97.8)	5233 (100)	115 (2.2)	0
/ɔ/	5182 (100)	5182 (100)	0	0
/u/	2212 (42.6)	5131 (98.8)	2981 (57.4)	62 (1.2)
Silence samples	62 (0.37)	84 (0.5)	16 787 (99.63)	16 765 (99.5)

Each column represents the samples in an identified class, whereas each row represents the samples in an actual class. The entry in the table is the number of identified samples; the number in parentheses is percentage of sample number.

larger than 0.37% in the normal subject. Post hoc tests showed that the confusion matrices of *b/a* method were significantly different across subjects ($P < 0.05$).

Reaction time test

Figure 4 shows the VIT and VTT of ANN-EL, video-EL, button-EL, and normal voice. For each subject, both the VIT and VTT were significantly different across voice sources (ANOVA, $P < 0.05$). For normal subjects, the VIT and VTT of ANN-EL were 311.8 and 409.8 milliseconds, respectively, which were significantly larger than normal voice and button-EL (post hoc, $P < 0.05$), but were significantly smaller than video-EL (post hoc, $P < 0.05$). For laryngectomized subject, the VIT of ANN-EL was 409.3 milliseconds, which was smaller than button-EL (469.5 milliseconds) and video-EL (575.6 milliseconds). Meanwhile, the VTT of ANN-EL in laryngectomee was 667.85 milliseconds, which was also smaller than the 833 milliseconds of video-EL but much larger than the 386.9 milliseconds of button-EL.

During experiments, it was observed that the precedence relationship between lips opening/closing and voice starting/stopping was closely related to on/off control of EL. Therefore, the relative time (RT) of lips opening/closing to voice starting/stopping was measured according to the video of lips movement and synchronous speech. For normal subject, the voice starting/stopping referred to the voice initiation/termination of normal speech production. But the normal speech was impossible to be produced in laryngectomee. Because the reaction time of button-EL for normal subjects was not significantly different from that of normal voice in this work and previous reports,[7,17] the voice initiation/termination of button-EL speech was regarded as the voice starting/stopping of normal speech. Figure 5 showed that the RT of all subjects were not significantly different across vowels (ANOVA, $P > 0.05$). At voice starting, the average RT of normal subjects was −7 milliseconds, which was significantly larger than −103.2 milliseconds of laryngectomee (*t* test, $P < 0.05$). But at voice stopping, the average RT of normal subjects was 114.9 milliseconds, which was significantly smaller than 205.8 milliseconds of laryngectomee (*t* test, $P < 0.05$).

For Chinese syllables, the consonant in word-terminal position can only be /n/ or /ŋ/, so only the influence of word-initial consonants is listed in Table 3. The VIT of ANN-EL were measured and compared with that of normal voice, which was also substituted by button-EL in the case of laryngectomee. The influences were summarized into three categories: "+" standing for significantly slower VIT than normal voice, "−" for significantly faster VIT, and "±" for equal VIT. The results in normal subjects were different across consonants, whereas the results in laryngectomee almost belonged to the same "−" category.

TABLE 2.
Confusion Matrix of Switch Identification for ANN and *b/a* Method in the Laryngectomee

	Phonation		Silence	
Testing Samples	*b/a*, N (%)	ANN, N (%)	*b/a*, N (%)	ANN, N (%)
Phonation samples				
/ɑ/	4740 (100)	4740 (100)	0	0
/i/	2276 (50.2)	4536 (100)	2260 (49.8)	0
/e/	4776 (98.5)	4848 (100)	72 (1.5)	0
/ɔ/	4772 (100)	4772 (100)	0	0
/u/	2428 (51.0)	4758 (100)	2330 (49.0)	0
Silence samples	1026 (7.2)	43 (0.3)	13 296 (92.8)	14 279 (99.7)

Each column represents the samples in an identified class, while each row represents the samples in an actual class. The entry in the table is the number of identified samples; the number in parentheses is percentage of sample number.

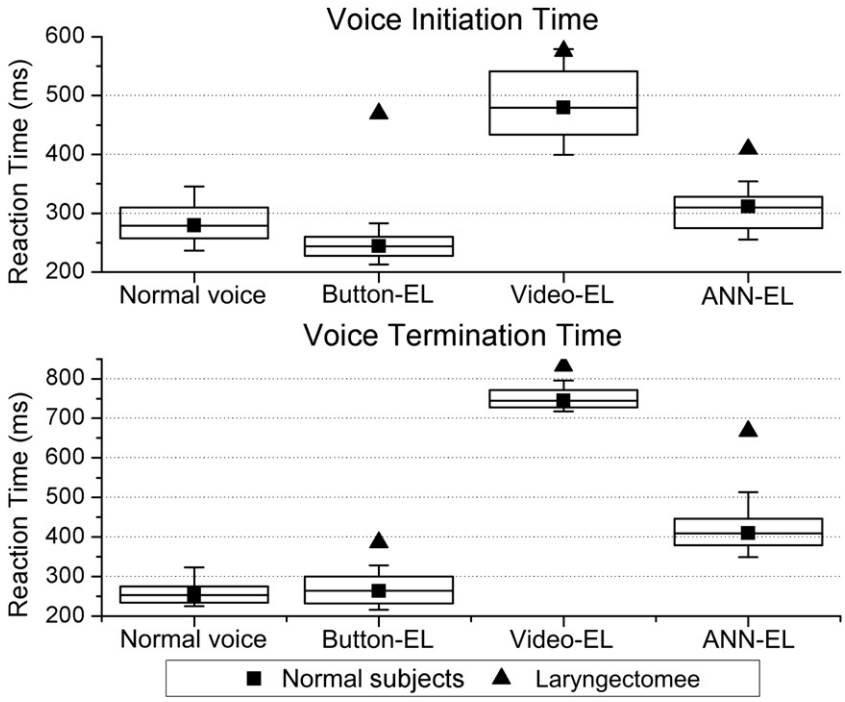

FIGURE 4. Voice initiation time and voice termination time measured for normal voice, button-EL, video-EL, and ANN-EL in normal subjects and laryngectomee. Box plots show the median, interquartile, and minimum-maximum values for four normal speakers; solid triangles show reaction times of the laryngectomized subject.

Fluency test

The results of speech fluency are listed in Table 4. The break errors of ANN-EL were 0.71% in normal subjects and 0.99% in laryngectomee, which were much less than that of video-EL. The voicing errors of ANN-EL were 6.25% in normal subjects and 12.5% in laryngectomee, which were also less than that of video-EL.

Intelligibility test

The average intelligibility scores of all subjects are shown in Figure 6. The intelligibility scores of ANN-EL were 31.94% for CV words and 46.07% for CVC words, respectively, which were significantly higher than that of video-EL and lower than that of button-EL (post hoc, $P < 0.05$). For normal subjects, the intelligibility scores of ANN-EL were 33.02% for CV words and 49.97% for CVC words, which were significantly higher than that (27.79% for CV words and 43.22% for CVC words) of laryngectomee (t test, $P < 0.05$). On the contrary, the intelligibility scores of button-EL and video-EL were not significantly different across subjects (t test, $P > 0.05$).

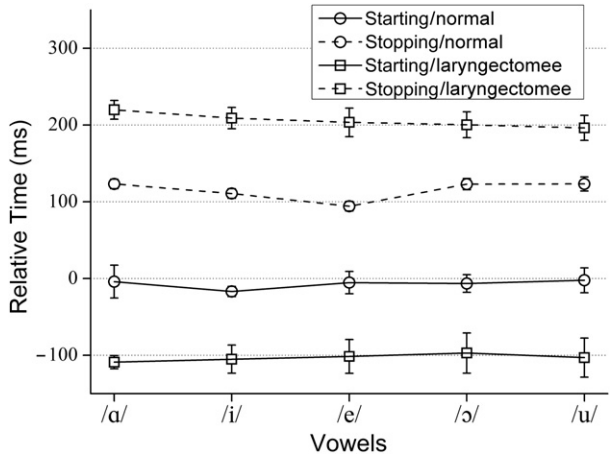

FIGURE 5. Relative time of lips opening/closing to voice starting/stopping across vowels. The normal speech is used as a reference of voice starting/stopping for normal subjects, whereas alaryngeal speech with button-EL is used for laryngectomee.

In addition, to study the influence of on/off control performance on intelligibility, initial and terminal misinterpretation errors were calculated. The initial misinterpretation error was defined as the percentage of incorrect response to word-initial consonants within all words. The terminal misinterpretation error was defined as the percentage of incorrect response to CV words with a nasal rhyme ending. The results are shown in Figure 7. For initial misinterpretation, the ANN-EL error of normal subjects was 38.7%, which was significantly lower than 63.21% of laryngectomee (t test, $P < 0.05$). In contrast, the initial misinterpretation errors of button-EL and video-EL were not significantly different across subjects. For terminal misinterpretation, the ANN-EL errors of normal subjects and laryngectomee were 20.76% and 26.27%, respectively, which were significantly lower than that of video-EL, but higher than that of button-EL (post hoc, $P < 0.05$).

DISCUSSION

Switch identification performance of ANN

The average total error of ANN (0.65%) in all subjects was significantly lower than that of *b/a* method (28.64%), which

TABLE 3.
The Influence of Word-Initial Consonants on Voice Initiation Time

Word-Initial Consonants	Normal Subjects	Laryngectomee
Lateral		
/l/	±	−
Nasal		
/m/	+	+
/n/	±	−
Fricative		
/f/	+	−
/h/	±	−
/sh/	±	−
Affricate		
/j/	±	−
/zh/	±	−
/ch/	±	−
Plosive		
/p/	+	+
/b/	+	+
/t/	−	−
/d/	−	−
/k/	−	−
/g/	−	−

The sign "+" means a slower VIT with the consonant in word-initial position than normal voice; the sign "−" means a faster VIT; the sign "±" means an equal VIT. For laryngectomee, alaryngeal speech with button-EL is substitute for normal voice as a reference.

indicated a more accurate switch identification of ANN method. All of these were mainly because of multiple parameters selection and nonlinear nature of ANN. For single parameter *b*/*a*, the clustering of phonation and silence were overlapped, so misidentification was inevitable. The strong evidence was high type I errors of /i/ and /u/ samples, which had a similar *b*/*a* value with silence samples and were hardly identified from silence using a linear threshold. For ANN method, four parameters separated phonation from silence in high-dimensional parameters space; meanwhile, nonlinear classification can identify phonation from silence as much as possible.

Another difference between ANN and *b*/*a* method was the individual differences of switch identification. For *b*/*a* method, the large standard deviation of error indicated significant differences from individual to individual, which mainly resulted from the threshold judgment. The threshold used in this article was an empirical value with 135% of baseline.[17] However, the baseline was set as parameter value of silence, which was different between individuals according to different lips. For instance, in Table 1, the thick lips of normal subject resulted in a large baseline and high threshold, which required more lip deformation to be identified as phonation. Therefore, the /i/ samples with small lip deformation were prone to be identified as silence, and silence samples were not easy to be misidentified. On the contrary, in Table 2, the thin lips of laryngectomee had a relative lower error of /i/ samples and higher error of silence samples. But ANN method can eliminate individual differences through ANN training to get accurate switch identification. Consequently, ANN method has more potential to control on/off of EL than *b*/*a* method.

On/off control performance of ANN-EL

The VIT and VTT are the most important parameters to evaluate on/off control performance. In this work, the VIT and VTT of normal voice were 279.5 and 253.9 milliseconds, respectively, which were approximately equal to the 274 and 240 milliseconds reported by Goldstein et al,[7] and the 282 and 269 milliseconds by Wan et al.[17] Meanwhile, the VIT (244.4 milliseconds) and VTT (263.9 milliseconds) of normal subjects using button-EL were also close to the measurements of Goldstein et al[7] and Wan et al.[17] Based on all of the extremely similar results with those reported by other researchers, it was concluded that our procedure of collecting VIT and VTT was valid and the results were comparable. In the case of laryngectomee, both VIT (469.5 milliseconds) and VTT (386.9 milliseconds) of button-EL were prominently larger than normal subjects, which might be explained by slower reaction resulting from the older age of the laryngectomee.

For normal subjects, the ANN-EL produced a slightly slower voice initiation than normal voice and button-EL. But in the case of laryngectomee, it was notable that the ANN-EL voice initiation was approximately 60 milliseconds faster than that of button-EL. This significant difference was closely related to the precedence relationship between lip-shape opening and voice starting. The RT results showed that the lips opening of normal subjects and voice starting were almost at the same time. But the laryngectomee opened lips much earlier than speech production, which was probably owing to the habits and mastery of using button-EL for a long time. Laryngectomized speaker was used to constructing the shape of vocal tract and lips before pressing the button. Therefore, taking the time of program processing (<50 milliseconds) into account, the voice initiation differences of ANN-EL from normal voice and button-EL can be well explained by the precedence relationship between lips opening and voice starting.

TABLE 4.
Break and Voicing Errors of Speech Produced by Video-EL and ANN-EL

	Normal Subjects		Laryngectomee	
EL	Break Errors (%)	Voicing Errors (%)	Break Errors (%)	Voicing Errors (%)
Video-EL	15.30	18.75	19.04	18.75
ANN-EL	0.71	6.25	0.99	12.50

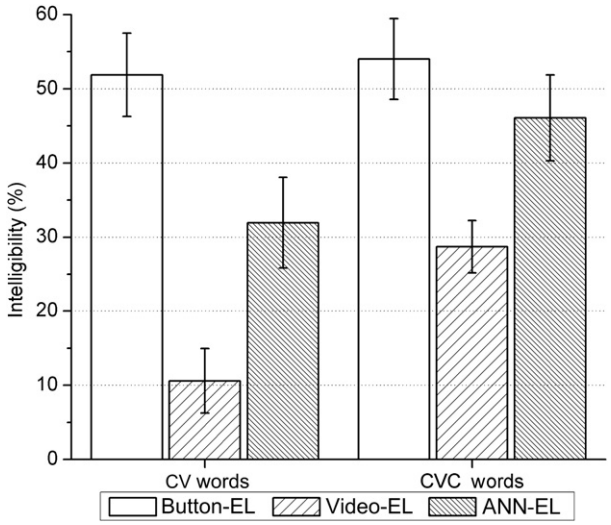

FIGURE 6. Intelligibility scores of CV/CVC words produced by button-EL, video-EL, and ANN-EL. The bars show the mean scores and standard deviations of word intelligibility for all listeners.

The VIT results showed that the video-EL voice initiation was slower than ANN-EL in all subjects, which was owing to the hysteresis of the *b*/*a* method. For video-EL, it took a little time for lips to deform over a fixed degree (the threshold) before EL switching on.[17] In contrast, ANN-EL was more sensitive to detect lips deformation because of nonlinear classification in multiparameter space, which hardly introduced any hysteresis. Besides, the large standard deviation of video-EL VIT also indicated the individual differences of hysteresis, yet which was not found in ANN-EL.

On the other hand, the VTT results showed that ANN-EL was not as fast as normal voice and button-EL in all subjects. The precedence relationship between lip closing and voice stopping also had an important impact on the performance of voice termination. As shown in Figure 5, each subject closed lips significantly later than voice stopping, which could be explained by actual process of speech production that the shape of vocal tract and lips was necessary to keep longer for a complete expression, and lip closing was not as fast as lip opening.[24] Furthermore, the ANN-EL voice termination was faster than ANN-EL in all subjects. Slow VTT of video-EL was mainly because of the 300-milliseconds off-delay, which was used to improve fluency through decreasing the unwanted breaks resulting from instantaneous drop below the threshold.[17] However, ANN-EL was more robust to noise than video-EL because of the accurate switch identification of ANN method. Subjects using ANN-EL were not only able to produce a faster VTT but also keep a good fluency, which will be discussed in the Speech Fluency section.

Although the results were obtained from different experimental subjects, the comparison with EMG-EL was possible and valuable based on the relative value to normal voice and button-EL in each experiment. For voice initiation, the EMG-EL was as fast as normal voice and button-EL, which indicated that ANN-EL voice initiation might not be slower than EMG-EL, especially in the laryngectomized subject. This might be explained by the fact that the neck trap muscle EMG precedes voice by 70 or 120 milliseconds,[25] which was approximately equal to the RT of lip opening before voicing started in the laryngectomee. For voice termination, the RT to button-EL showed that ANN-EL VTT might be much smaller than EMG-EL in laryngectomized subject. This finding can be explained by the fact that lip closing is an active process, whereas EMG-EL lacks a corresponding active mechanism for voice termination, and postphonatory activity of laryngectomized subject's trap muscles lingers EMG-EL phonation.[7,26]

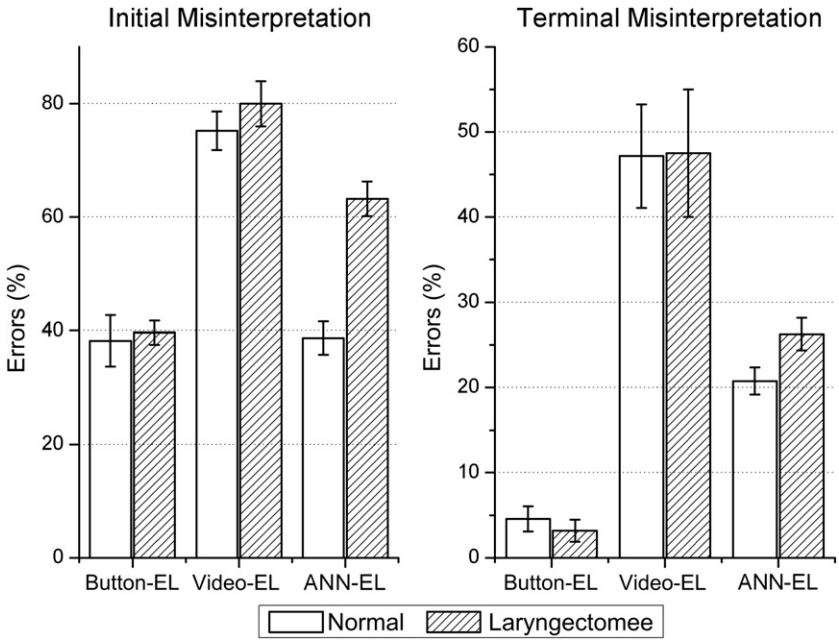

FIGURE 7. Initial and terminal misinterpretation errors of words produced by button-EL, video-EL, and ANN-EL in normal subjects and laryngectomee. The bars show the mean errors and standard deviations for all listeners.

All the above-mentioned experiments were discussed in the case of single vowel, which was uncommon in daily communication. The consonant is an indispensable phoneme to express a meaningful word, and also has some lip-shape pattern during phonation, which has an impact on the on/off control of EL. The results showed that the plosive consonants /t/, /d/, /k/, and g/ produced a faster voice initiation than normal voice in all subjects. Because it should block vocal tract and shape lips before a sudden release of the compressed air, these plosive consonants are all with open mouth. In contrast, the plosive consonants /p/ and /b/ and nasal consonant /m/ with initial closed mouth were identified to be silence samples, supposed to be phonation samples, and led to a slower voice initiation. For other consonants, there were no significant differences between ANN-EL and normal voice in normal subjects, whereas most of the consonants made a faster voice initiation in laryngectomee. This observation might relate to much earlier lip opening than voice starting in laryngectomee. In the case of word-terminal consonants, the nasal rhymes /n/ and /ng with closed mouth might switch ANN-EL off a little earlier, which however could not completely offset the delay of lip closing to voice stopping. So the voice termination of ANN-EL was closer to normal voice, but still slower.

Speech fluency

The fluency results showed that subjects produced a fluently long sentence using ANN-EL. First, the less break error of ANN-EL indicated a better continuity of alaryngeal speech than video-EL. The phonemes with closed mouth, such as /i/, were easily identified as silence in video-EL, which caused more breaks in the continuous speech. Second, the less voicing error showed that enough pause intervals could be produced by ANN-EL to distinguish two sentences. The more voicing error of video-EL was owing to the 300-milliseconds off-delay, which decreased the pause interval.

Furthermore, the 300-milliseconds off-delay of video-EL was set to improve the continuity of speech, consequentially, which caused a slower voice termination and more voicing error.[17] However, ANN-EL without off-delay can make a better fluency, which can be explained by the following reasons. First, lip closing is slower than voice stopping, which ensures the continuity of speech on its own. Second, the shape of lips deforms continuously and hardly returns to that of silence during continuous speech. Therefore, off-delay is not necessary for ANN-EL. In addition, the influence of speaking rate is not discussed in this article, which however is an important influential factor for both ANN-EL and video-EL.[17] Although the speech fluency of ANN-EL was satisfied at normal speed, it was clear that the current frame rate of video was not enough for high speaking rate, which might affect the lip shape of phonation and reduce on/off control performance of ANN-EL.[27]

Speech intelligibility

The intelligibility results showed that there were significant differences across voice sources, which indicated a strong impact of on/off control method on word intelligibility. For both CV and CVC word, subjects using ANN-EL produced a more intelligible word than video-EL, but less intelligible word than button-EL, which was identical with the ranking of on/off control performance. Therefore, the intelligibility differences across voice sources can be explained by differences of on/off control performance, which was mainly reflected in two aspects: misinterpretations of word-initial consonants and nasal rhymes ending.

The results showed that normal subjects produced a more intelligible word than laryngectomee using ANN-EL, which was related to the more initial misinterpretation error of laryngectomee than normal subjects. The more initial misinterpretation in laryngectomee was probably owing to the faster voice initiation of ANN-EL resulting from the much earlier lip opening. Besides, the ANN-EL initial misinterpretation of normal subjects was equal to that of button-EL, which was probably attributed to the synchronous lip opening with voice starting in normal subjects. This indicated that the synchronism of on/off control and voice starting/stopping was important for on/off control performance and intelligibility. Thus, the slowest voice initiation of video-EL led to the highest initial misinterpretation errors and least intelligibility scores.

On the other hand, the intelligibility of CVC words was a significantly higher than CV words for ANN-EL and video-EL (t test, $P < 0.05$) owing to the slower off-control than voice stopping. The delayed voice termination of EL leads to an extended ending, which is always misinterpreted as nasal rhymes /n/ or /ng/ followed by CV words. The terminal misinterpretation results showed that the nasal rhyme ending was more likely to occur in CV words produced by video-EL and easily perceptible with longer VTT. The terminal misinterpretations of ANN-EL were lower than video-EL, but higher than button-EL, which made a contribution to the lower intelligibility of ANN-EL than button-EL. As for CVC words, the influence of delayed voice termination was smaller. This is mainly because of the fact that Chinese CVC syllables can only end with nasal rhymes /n/ or /ng/. But word-terminal consonants in CVC words were still confusing owing to nasal rhymes ending, so the intelligibility of ANN-EL was still lower than button-EL.

Therefore, the lower intelligibility of ANN-EL than button-EL is mainly owing to the asynchronism of on/off control and voice starting/stopping, which might be improved through training and practice. In addition, the perception in our testing confined to a single word. When producing a sentence, the intelligibility of words in the sentence might be higher with the fluent voice source of ANN-EL.[17]

CONCLUSION

In this work, ANN was proposed to control the on/off mechanism of EL based on visual information of lips. The switch identification of ANN was proved to be more accurate than *b*/*a* method. The ANN-EL system was developed to implement the real-time on/off control. On/off control performance, speech fluency, and word intelligibility of ANN-EL were evaluated and compared with other voice sources. For on/off control performance, ANN-EL was as fast as normal voice and button-EL in voice initiation, but slower in voice termination.

Furthermore, the intelligibility of ANN-EL was not as high as button-EL owing to the worse performance of on/off control. However, ANN-EL could provide a more fluent and intelligible speech than video-EL. The current ANN-EL system based on a personal computer is not suitable for direct application in daily communication. Now a portable ANN-EL device is being developed, and our future work will focus on the training effect and the ANN-EL performance in daily life.

Acknowledgments

The authors would like to express special appreciation to Ye Wenyuan and Rui Baozhen for their works in the experiments. This work was supported by the National Natural Science Foundation of China under Grants 10874137, 11274250, and 61271087.

REFERENCES

1. Liu HJ, Ng ML. Electrolarynx in voice rehabilitation. *Auris Nasus Larynx*. 2007;34:327–332.
2. Morris HL, Smith AE, Van Dmark DR, Maves MD. Communication status following laryngectomy: the Iowa experience 1984-1987. *Ann Otol Rhinol Laryngol*. 1992;101:503–510.
3. Boone DR, McFarlanc SC, VonBerg SL, Zraick RI. *The Voice and Voice Therapy*. Englewood Cliffs, NJ: Prentice Hall; 1988.
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. Knorr SG, Zwitman DH. 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. 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.
8. 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.
9. Heaton JT, Goldstein EA, Kobler JB, 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.
10. 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.
11. Dupont S, Luettin J. Audio-visual speech modeling for continuous speech recognition. *IEEE Trans Multimed*. 2000;2:141–151.
12. Tomlinson MJ, Russell MJ, Brooke NM. Integrating audio and visual information to provide highly robust speech recognition. *Proc IEEE Int Conf Acoust Speech Signal Process*. 1996;2:821–824.
13. Sanderson C, Paliwal K. *Information fusion and person verification using speech and face information*: IDIAP Research Report, Martigny, Switzerland; 2002:02–33.
14. Wark T, Sridharan S. A syntactic approach to automatic lip feature extraction for speaker identification. *Proc IEEE Int Conf Acoust Speech Signal Process*. 1998;6:3693–3696.
15. Evitts PM, Portugal L, Dine AV, Holler A. Effects of audio-visual information on the intelligibility of alaryngeal speech. *J Commun Disord*. 2010;43: 92–104.
16. Evitts PM, Dine AV, Holler A. Effects of audio-visual information and mode of speech on listener perceptions of alaryngeal speakers. *Int J Speech Lang Pathol*. 2009;11:450–460.
17. Wan CY, Wu L, Wu HX, Wang SP, Wan MX. Assessment of a method for the automatic on/off control of an electrolarynx via lip deformation. *J Voice*. 2012;26:674.e21–674.e30.
18. Bishop CM. *Neural Networks for Pattern Recognition*. Oxford, UK: Oxford University Press; 1995.
19. Yao H, Lu Y, Gao W. Lip-movement features extraction and recognition based on chroma analysis. *Acta Electronica Sinica*. 2002;30: 168–172.
20. Otsu N. A threshold selection method from gray-level histograms. *IEEE Trans Syst Man Cybern*. 1979;9:62–66.
21. Cybenko G. Approximation by superpositions of a sigmoidal function. *Math Control Signals Syst*. 1989;2:303–314.
22. Liu HJ, Wan MX, Wang SP. Features of listeners affecting the perceptions of mandarin electrolaryngeal speech. *Folia Phoniatr Logop*. 2005;57: 9–19.
23. Sammut C, Webb GI. *Encyclopedia of Machine Learning*. 1st ed. New York, NY: Springer; 2011.
24. Goffman L, Smith A. Development and phonetic differentiation of speech movement patterns. *J Exp Psychol Hum Percept Perform*. 1999;25: 649–660.
25. Atkinson J. Correlation analysis of the physiological factors controlling fundamental voice frequency. *J Acoust Soc Am*. 1978;63:211–222.
26. Hillel 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.
27. 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.