Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 57 (2014) 39–49

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Evaluation of a method for vowel-specific voice source control of an electrolarynx using visual information

Liang Wu, Congying Wan, Ke Xiao, Supin Wang *, Mingxi Wan *

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, PR China

Received 6 March 2013; received in revised form 11 September 2013; accepted 14 September 2013
Available online 21 September 2013

Abstract

The electrolarynx (EL) is a widely used device for alaryngeal communication, but the low quality seriously reduces the intelligibility of EL speech. To improve EL speech quality, a vowel-specific voice source based on visual information of lip shape and movements and artificial neural network (ANN) is implemented into an experimental EL (SGVS-EL) system in real time. Five volunteers (one laryngectomee and four normal speakers) participated in the experimental evaluation of the method and SGVS-EL system. Using ANN participants were able to perform high vowel precision with identification rates of >90% after the training. The results of voicing control indicated that all subjects using SGVS-EL could achieve good vowel control performance in real time, but still control errors frequently occurred at the voice initiation period. However, the control errors had no significantly impact on the perception of SGVS-EL speech. Intelligibility evaluation demonstrated that both the vowels and words produced using the SGVS-EL were more intelligible than vowels spoken with a commercial EL (by 30%) or words (by 18%), respectively. Using a controlled vowel-specific voice source was a feasible and effective way to improve EL speech quality with more intelligible words.

Keywords: Electrolarynx; Intelligibility evaluation; Visual information; Vowel-specific voice source

1. Introduction

Standard esophageal (SE) speech, tracheo-esophageal (TE) speech, and electrolarynx (EL) speech are three main ways of voice rehabilitation for laryngectomized patients (Pawar et al., 2008; Carr et al., 2000). Due to easy learning and no additional surgery, more than half of laryngectomized patients use an EL as their primary way of daily communication (Carr et al., 2000; Clements et al., 1997; Morris et al., 1992). Despite widespread use of EL, the unnatural quality is still a serious drawback that limits the understanding of EL speech (Hillman et al., 1998; Meltzner et al., 2005). EL voice source plays an important role in EL speech production. Multiple studies have reported that the improper characteristics of EL voice source was one of the main reasons to reduce the EL speech quality (Weiss et al., 1979; Qi and Weinberg, 1991; Meltzner, 2003). To improve EL speech quality, we have previously proposed a supra-glottal voice source provided through a PC-based experimental EL (SGVS-EL) system Wu et al., 2013. The supra-glottal voice source was a compensated glottal source with vocal tract characteristics according to different vowels. The acoustic properties of SGVS-EL speech were measured and analyzed comparing with normal speech and commercial EL speech. The results indicated that the supra-glottal voice source was feasible and effective to improve the acoustic quality of EL speech by enhancing low-frequency energy,

* Corresponding authors. Address: No. 28, Xianning West Road, Xi'an, Shaanxi 710049, PR China. Tel.: +86 29 82667924; fax: +86 29 83237910 (M. Wan).
E-mail addresses: spwang@mail.xjtu.edu.cn (S. Wang), mxwan@mail.xjtu.edu.cn (M. Wan).

0167-6393/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2013.09.006

correcting the shifted formants to normal range, and eliminating the visible spectral zeros.

The previous SGVS-EL can only provide one kind of supra-glottal voice source for each phonation. Because the supra-glottal voice source is vowel-dependent, real time identification of vowels is essential to control synthesis and output of EL voice source for continuous speech. Many methods have been reported on EL voicing control. Uemi used expiration pressure to control the intonation of EL speech (Uemi et al., 1995). Takahashi designed a denture-based intra-oral EL, using intra-oral pressure to detect utterance of voiceless consonants and control the EL voicing (Takahashi et al., 2008). In addition, Goldstein and Stepp used electromyography (EMG) to control the onset, offset, and pitch of EL speech (Goldstein et al., 2004, 2007; Stepp et al., 2009). All of these methods were proved to be effective, but none of these physiological signals cannot be used to detect the vowel types during phonation in real time.

Visual information, especially the shape information of lips, has been widely applied in speech recognition (Dupon and Luettin, 2000; Neti et al., 2000) and speech perception (Massaro and Cohen, 1983; Evitts et al., 2010), because it reflects the process of speech production and contains speech-related cues. Lip-reading is such an useful technique of understanding speech by interpreting visual movements (Summerfield, 1992). Therefore, visual information is a potential signal for EL voicing control. We have previously demonstrated that it is feasible to control onset and offset of EL speech using lips deformation (Wan et al., 2012; Wu et al., 2013).

In this article, our purposes mainly focused on two aspects: firstly, we proposed a real time method of vowel identification based on visual lips information to control the vowel-specific voice source. The previous SGVS-EL system was updated with implementation of the control method, and the performance of SGVS-EL system for voice source control was assessed. Secondly, previous work only analyzed the acoustic properties of EL speech produced with supra-glottal voice source (Wu et al., 2013). Therefore, the intelligibility of EL speech produced by the new SGVS-EL was evaluated and compared with commercial EL speech, to confirm further feasibility of the controlled vowel-specific voice source to improve EL speech quality.

2. Experimental electrolarynx (SGVS-EL) system

A schematic diagram of the SGVS-EL system is shown in Fig. 1. This system utilizes a camera to capture video signal of lips and processes the visual information to identify phonation vowel. The vowel type is then used to control the synthesis of supra-glottal voice source. The SGVS-EL system contained three main modules: control module, supra-glottal voice source synthesis module, and output module.

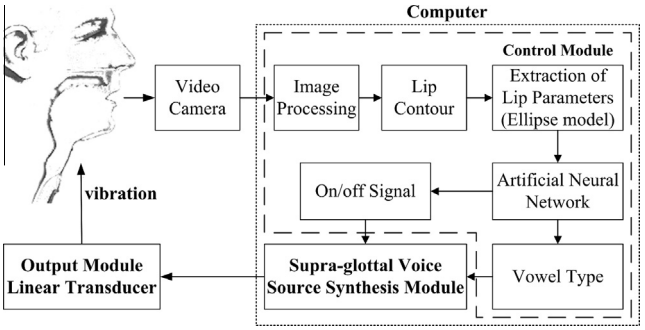

Fig. 1. Schematic diagram of the SGVS-EL system. The dash line box represents the control module. The procedure in dotted line box is processed through a program in a desktop computer.

2.1. Control module

The control module is to realize the vowel identification based on visual information of lips. As shown in Fig. 1, the facial image captured by a camera is processed to obtain the lip contour through background preprocessing and lip segmentation (Wan et al., 2012; Wu et al., 2013). Then, the lip contour is fitted by an ellipse model with two parameters, namely semi-major axis (a) and semi-minor axis (b). Considering the nonlinear map between the vowel and corresponding lip shape, a two-layer feed-forward artificial neural network (ANN) is used for vowel identification.

The pair of parameters (a, b) is transformed to a new vector as input of the ANN. The input vector consists of normalized semi-major axis (a/a_0), normalized semi-minor axis (b/b_0), ratio of b to a (b/a), and normalized area of the ellipse (ab/a_0b_0). The normalized factors (a_0, b_0) are set during system initialization, representing the lip parameters of silence. The normalized parameters can reduce the influence of head movement. Twenty neurons are selected empirically in hidden layer. For each neuron, the net function is a weighted linear combination, and the activation function is a hyperbolic tangent activation function. In output layer, six neurons represent five different vowels (/ɑ/, /i/, /e/, /ɔ/, and /u/) and silence, respectively. As a result, the ANN output is the control signal for synthesis of supra-glottal voice source.

2.2. Supra-glottal voice source synthesis module

The supra-glottal voice source and its synthetic procedures are described in a separate article (Meltzner, 2003). Briefly, the glottal voice source is compensated with vocal tract characteristics according to different vowels, resulting in vowel-dependent supra-glottal voice source. In this system, the synthesis of supra-glottal voice source is regulated by the output signals of the control module in two ways: first, the vowel type of ANN output determines the vocal tract parameters selected for synthesis; second, the silence of ANN output determines the cessation of synthetic process. Therefore, the SGVS-EL is able to control both the

Fig. 2. The experimental electrolarynx (SGVS-EL) system. (a) Photo of system components. (b) The SGVS-EL is used by an individual.

vowel type and starting/stopping of the supra-glottal voice source at the same time.

2.3. Implementation of SGVS-EL system

As shown in Fig. 2(a), the SGVS-EL system consists of a modified microphone headset (Danyin DT-2699, Guangzhou, China), a desktop computer (KaiTian A6000, Lenovo, Beijing, China), a driving circuit, and a wearable linear EL vibrator. A camera (Lenovo, Beijing, China) is fixed to the modified microphone headset via an adjustable metal support, maintaining a relatively constant position to the lips, which can avoid the influence of head movement on video capture during speaking. The video of lips is captured by the camera and send to the computer via USB port. A PC-based program is developed to implement the methods of the control module and supra-glottal voice source synthesis module, and output the digital waveforms of supra-glottal voice source to the driving circuit via a parallel port. The driving circuit and EL vibrator compose the output module. The driving circuit performs D–A conversion and power amplification to drive the linear vibrator, which is made of a mini-speaker (Somic SN-401, Guangzhou, China) with a 3 cm diameter. The EL vibrator is able to provide an sufficiently linear voice source with acceptable distortion (Wu et al., 2013). Fig. 2(b) shows an individual using the SGVS-EL. The camera is about 10 cm in front of the lips with the microphone located approximately 5 cm from the mouth. The EL vibrator is hold by a hand to guarantee a high coupling of vibrator with neck tissue. Finally, the SGVS-EL can produce a speech with 60 ± 3 dB SPL intensity.

3. Experiments

3.1. Participants

Five subjects participated in this study. Four normal participants (N1–N4) were recruited as control group, including 2 male (ages 25 and 26 years) and 2 female (ages 22 and 24 years) with no history of voice disorders. Those participants were moderately proficient at using commercial EL. The remaining participant (L) was a male laryngectomee (age 76 year) who had 2 years' experience of using commercial EL after a total laryngectomy surgery.

Given that the surgery would not influence the lip shape of phonation (Wu et al., 2013), normal healthy participants were studied for two purposes: (a) to validate the control method and the experimental procedure for training and (b) to use the normal participant outcomes as a proxy for the potential performance of the laryngectomized participant.

3.2. Training protocol

The experimental protocol involved two alternative trainings (ANN training and phonation training) for 9 consecutive days. The ANN training was performed in odd days (day 1, 3, 5, 7, and 9), to collect samples for training and testing network, and to compute the clustering of different vowels for guiding phonation training. The phonation training was performed in even days (day 2, 4, 6, and 8), to instruct participants to form different and stable lip shapes of different vowels for improving performance of vowel identification.

At the beginning of day 1, an investigator reviewed with the participant the basic information, including the relationship between the lip shape and the vowel identification, and the training procedures. Then, the ANN training was performed to acquire the intrinsic lip shape pattern of different vowels without the phonation training. During the process, the participant was instructed to pronounce sustained vowels (/ɑ/, /i/, /e/, /ɔ/, and /u/) following few seconds of silence. The normal subjects spoke with normal voice while the laryngectomized subject with commercial EL (Servox digital, Servona, Germany). Meanwhile, the lip parameters (a and b) were extracted and recorded synchronously with the speech. According to the speech signal, the parameters were manually selected as different types of samples and transformed into pairs of inputs and outputs,

ruling out the situation as smile and yawn. Finally, all selected data were randomly mixed and equally divided into two group for ANN training and testing. The scaled conjugate gradient back propagation algorithm was used to train the network. In addition, all selected data were used to compute the clustering distribution of different vowels in the space of ANN input parameters.

On day 2, the phonation training contained five guided phonation sessions and a final practice session. During the guided phonation session, the participant repeated to pronounce one sustained vowel while receiving the feedback of vowel identification and instruction on how to improve performance. The guiding strategies were established, depending on the relationship between the lip shape and clustering distribution of different vowels. The instructions guided the participant to change and keep their lip shape to a target, which guaranteed the biggest differences from other vowels and obtained the best vowel identification. The target of each vowel was determined in the ANN training the day before. Each guided phonation session contained 15 min of focused practice on only one vowel. During the 30 min of final practice session, the participant was asked to pronounce any vowel randomly just with the feedback, which entailed participants practicing and enhancing the lip shape pattern in a self-directed manner.

All participants underwent the 9-day protocol. The training was firstly performed by participant N1–N4 to validate the procedure. After completing the training protocol, the ANN with best vowel identification for each participant was selected for the real time control of SGVS-EL system.

3.3. Materials and recording

Two sets of materials were used for the two speech tasks of the voicing control and intelligibility. One set contained five single vowels (/ɑ/, /i/, /e/, /ɔ/, and /u/). The other set was 70 monosyllabic words with consonant-vowel-consonant (CVC) structure as listed in Table 1, containing 5 vowels and 14 commonly used consonants in both word-initial and word-final positions.

Both sets were used to assess the voicing control of SGVS-EL. Firstly, the normal subjects (N1–N4) and laryngectomee (L) were asked to pronounce each sustained vowel with normal voice and commercial EL, respectively. Each trial lasted 2–4 s and repeated 10 times. Meanwhile, the SGVS-EL system was used to identify the vowel and control the output of EL voice source. The speech signals were recorded synchronously with the voice source signals of SGVS-EL. Secondly, the CVC words were used to estimate the influence of consonants on the voicing control.

Only the word set was used to evaluate the intelligibility of SGVS-EL speech. Each word was placed in a carrier sentence, "Write ___ again." to provide a consistent and more natural speech context. All participants were asked to read the materials with SGVS-EL, commercial EL, and normal voice (only the normal subjects), respectively, and the speech signals were recorded.

Table 1
The monosyllabic words with consonant-vowel-consonant structure.

/lɑ :d/	/li:ʃ /	/leʧ /	/lɔ :d/	/lu:m/
/mɑ :ʧ /	/mi:n/	/meɵ /	/mɔ :m/	/mu:n/
/nɑ :d/	/ni:k/	/neʃ /	/nɔ :ɵ /	/nu:p/
/fɑ :m/	/fi:t/	/fem/	/fɔ :k/	/fu:d/
/sɑ :k/	/si:b/	/set/	/sɔ :b/	/su:n/
/ʃ ɑ :p/	/ʃ i:ɵ /	/ʃ ed/	/ʃ ɔ :t/	/ʃ u:t/
/hɑ :ʃ /	/hi:t/	/hep/	/hɔ :k/	/hu:ʃ /
/pɑ :t/	/pi:k/	/pen/	/pɔ :ʃ /	/pu:ʧ /
/bɑ :ɵ /	/bi:d/	/bed/	/bɔ :n/	/bu:t/
/tɑ :n/	/ti:ʧ /	/ten/	/tɔ :ʧ /	/tu:ɵ /
/dɑ :k/	/di:m/	/deb/	/dɔ :t/	/du:b/
/kɑ :n/	/ki:p/	/kep/	/kɔ :n/	/ku:k/
/ɡ ɑ :b/	/ɡ i:n/	/ɡ et/	/ɡ ɔ :d/	/ɡ u:d/
/ʧ ɑ :t/	/ʧ i:p/	/ʧ ek/	/ʧ ɔ :p/	/ʧ u:k/

All the recordings were carried out in a soundproof room. Speech signals were collected using a data acquisition system (BioPac MP150, America) with a dynamic microphone (Salar M9, Guangzhou, China) mounted 10 cm away from the mouth. The acoustic data were digitized at a 44.1 kHz sampling rate with 16-bit quantization.

3.4. Data analysis

3.4.1. Vowel identification of ANN

The vowel identification of ANN was scored by the identification results of network. For each time of ANN training, the overall identification accuracy of the trained ANN was tested to assess the identification performance and the phonation training effects. After the 9-day training, the final trained ANN for each participant was tested and evaluated by two types of errors. Type I error was scored by false positive rate, defined as the percentage of misidentified samples within the identification results of each vowel. Type II error was scored by false negative rate, defined as the percentage of misidentified samples within the testing samples of each vowel.

3.4.2. SGVS-EL voicing control

The real time voicing control of SGVS-EL system was assessed by comparison of the SGVS-EL voice sources with the synchronous speech. Given that the on/off control performance had been studied in previous report (Wu et al.,

2013), this work mainly focused on the voice source control. For single vowel phonation, the control errors were calculated by the percentage of incorrect vowel type of the output voice source in different time periods of speech production, including initiation, duration, and termination. The time periods were defined according to the speech signals. The initiation period referred to the first 500 ms after the start point of speech signal (voice initiation), whereas the termination period referred to the last 500 ms before the end point of speech signal (voice termination). And the duration period was between the initiation period and termination period. For CVC words, the control errors were also measured and compared with those of single vowel, to estimate the influence of the consonants on voicing control. The 500 ms were chosen to make a complete vowel initiation and termination. Because in the case of CVC word, the sum of consonant duration and voice onset time was less than 400 ms in general (Umeda, 1977; Cho and Ladefoged, 1999).

3.4.3. Intelligibility

Intelligibility was measured by scoring target words via listener comprehension. Twelve young listeners participated in this task, including 7 males and 5 females aged from 22 to 28 years (averaging 25.2 years). All listeners had no reports of any hearing and language disorders. There were total 14 sets of 70 sentences with target words recorded, which were played to the listeners through headphones in a quiet room. To avoid learning effect, the order of sentences was set randomly. Listeners were instructed to transcribe the monosyllabic word heard using broad phonetic transcription, listening to the sentence as many times as needed. The average percentage of correct responses for all listeners was calculated as the intelligibility score. Twenty percent of judged trials were repeated by all listeners, yielding a high intrajudge reliability (94%) and interjudge reliability (85%).

4. Results and discussion

4.1. ANN identification performance

The overall identification accuracy of each trained ANN in the training protocol is shown in Fig. 3 for all participants. At the beginning of the protocol, the initial identification rates of all subjects were larger than 40%, especially 70% for subject N1. This result indicated that there was certain lip shape pattern during phonation of different vowels for each subject before the phonation training. However, the standard deviation of each subject was larger than 3%, which indicated that the intrinsic lip shape pattern was not stable. Even so, the intrinsic lip shape pattern was still able to provide sufficient initial condition and guidance for the phonation training.

As the training went on, the identification accuracy of each subject was gradually increased by about 30% to 50%. Meanwhile, the standard deviation of each subject

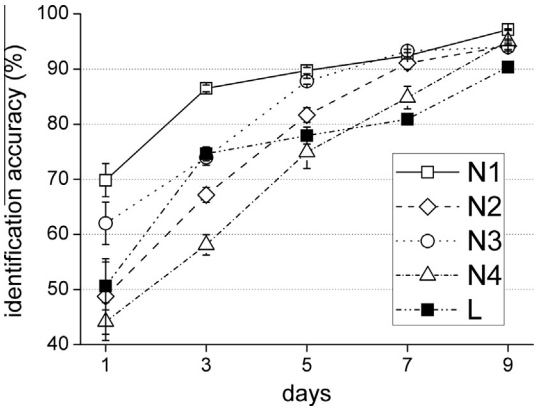

Fig. 3. Overall identification accuracy of the trained ANN in the training protocol for all subjects.

was reduced to fewer than 1%. These outcomes demonstrated that the strategy of phonation training was feasible and effective to form a stable pattern of lip shape based on the intrinsic one. And this lip pattern was able to improve the identification performance of ANN by maximizing the lip shape differences of different vowels. When the training was finished, the ANN identification accuracies of all participants were larger than 90%. However, the identification accuracy was higher, the improvement was harder. As the phonation training continues, the average increment of the identification accuracy for all subjects was 17.0%, 10.34%, 6.09%, 5.66%, respectively, showing a gradually decrease trend. Specially for the subject N1, the improvement was obviously difficult (less than 3%) when the identification accuracy was over 85%. This was mainly due to the method limitations, that the simplified model of lip shape and limited parameters could not totally represent the characteristics of lip shape and absolutely distinguish one vowel from others.

In the whole experiment, the initial identification accuracy and training effect of each subject were different from others, which implied potential differences from individual to individual. To reveal the differences of ANN identification, the two identification errors of final trained network were shown in Fig. 4 for two normal subjects (N1 and N2) and laryngectomized subject (L). First, the results of one subject showed different identification errors across the vowels. For subject N2, the two errors of vowel /ɑ/ and /e/ (about 15–20%) were much larger than that of vowel /i/, /ɔ/, and /u/ (lower than 5%). Because the identification depended on the dissimilarity of lip shape, the larger errors of vowel /ɑ/ and /e/ were actually owing to the similar lip shape between themselves. Second, the results of different subjects demonstrated significant difference across the individuals. The errors of subject N1 mainly occurred in the vowel /e/ and /u/, whereas in the vowel /ɑ/ and /e/ for subject N2, and among vowel /ɑ/, /i/, and /e/ for laryngectomee (L). In addition, the maximum error of subject N1 was 7.2%, which was much lower than that of subject N2 (18.2%) and L (23.4%). Considering the identi-

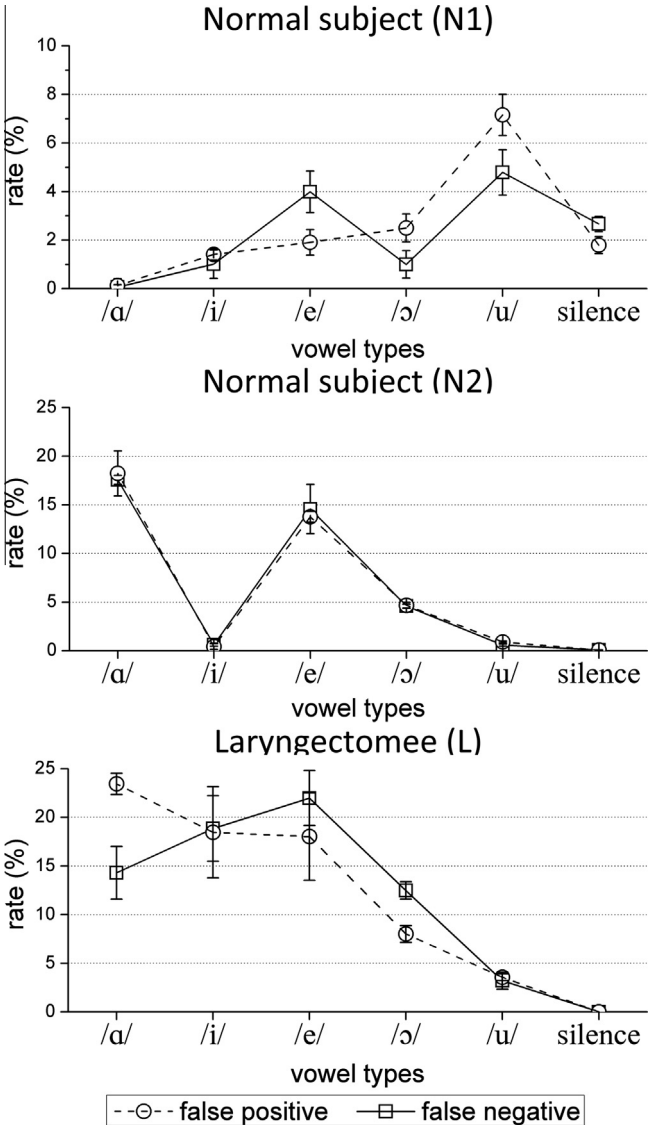

Fig. 4. The false positive rate and false negative rate of ANN identification for two normal subjects (N1 and N2), and laryngectomee (L).

fication criterion of the method, the individual difference might be due to the different lips and its patterns during phonation. Therefore, both physiological characteristics and phonation features of the lips had an important impact on the ANN identification. Even so, all participants could acquire high performance of ANN identification and show great potential for SGVS-EL voicing control.

4.2. SGVS-EL control performance

4.2.1. Single vowel control

The real time vowel identification and voice source control of SGVS-EL are shown in Fig. 5 for phonation of sustained vowel. The results showed that the real time control accuracy was as high as the identification accuracy of ANN identification. Even in the duration period of vowel phonation, the control accuracy was higher than 95%. This outcome indicated that the control method based on visual information and ANN was feasible to control SGVS-EL voice source in real time.

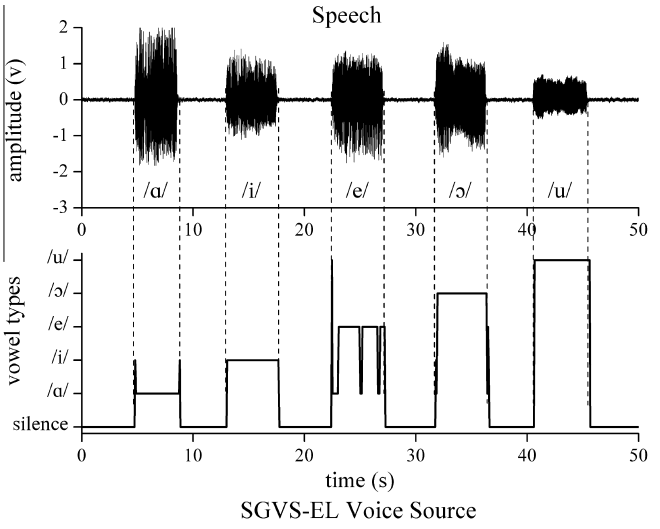

Fig. 5. Real time identification and voice source control of SGVS-EL during phonation of sustained vowels.

However, the control errors still happened more frequently at the transition time between two different phonation states. For instance, vowel /i/ was a prone error during the transition from silence sate to vowel /ɑ/ and the inverse progress. To quantitatively evaluate the control performance of SGVS-EL, the control errors in different time periods are shown in Fig. 6. Graph (a) showed that the average control errors of initiation period were 75.2% for normal subjects and 14% for laryngectomee, which were significantly larger than that of duration period (7.2% for normal subjects and 1.3% for laryngectomee). This result was mainly due to the larger difference of the lip shape from the target vowel at initiation period than duration period. In addition, the average control error of the laryngectomee was smaller than that of the normal subjects, especially in initiation period. This outcome might be due to the precedence relationship between lips opening and voice initiation. The previous study reported that normal subject opened lips as fast as voice initiation, but laryngectomee opened lips much earlier than voice initiation (Wan et al., 2012). For this reason, the beginning of changing progress in laryngectomee was earlier than normal subject, thus, the lip shape of laryngectomee was closer to target vowel than normal subject at the voice initiation, and the progress also ended earlier in laryngectomee. These two aspects reduced the misidentification probability and control errors in the case of laryngectomee. This conclusion was supported by the time-varying control errors of different subjects as shown in Fig. 6(b). On the one hand, at the voice initiation, the control errors of the laryngectomee was only 14%, which was much smaller than 74.8% of

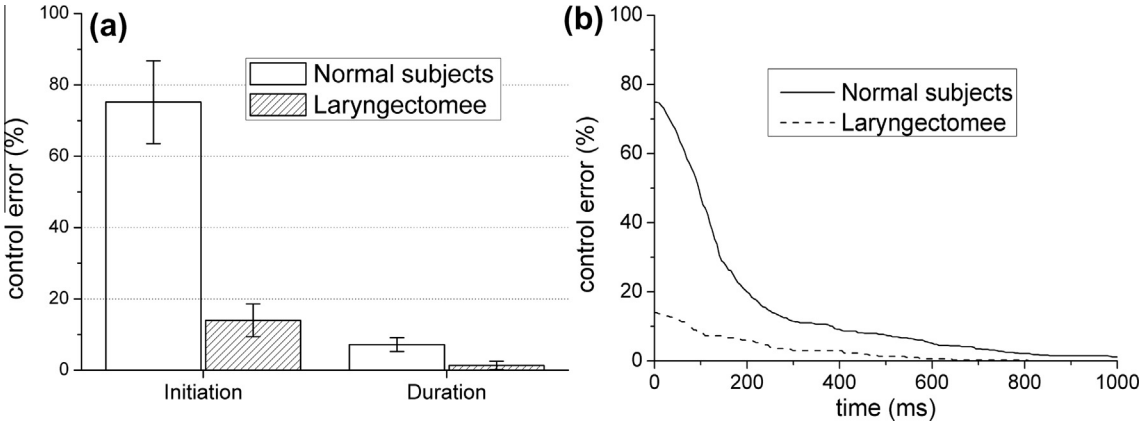

Fig. 6. Control errors in the single vowel production. (a) Control errors in initiation period (0–500 ms) and duration period (>500 ms). (b) Time-varying control errors of the normal subjects and laryngectomee.

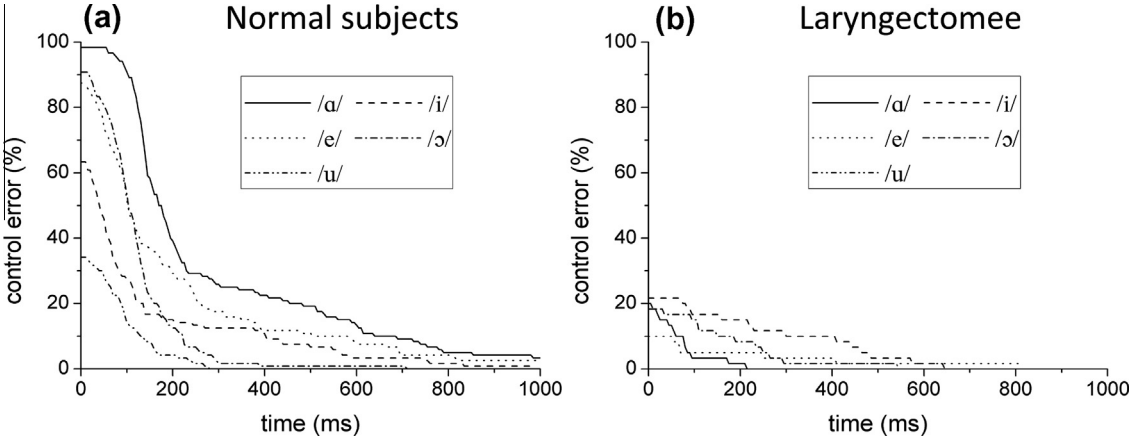

Fig. 7. Time-varying control errors of different vowels for (a) the normal subjects and (b) the laryngectomee.

the normal subjects. On the other hand, it took 230 ms for the laryngectomee to reduce the control error below 5%, while 600 ms for the normal subjects. Therefore, the control performance was influenced by the phonation habit, and the laryngectomized subject might acquire a better control of SGVS-EL voice source.

Furthermore, to reveal the differences across vowels, the time-varying control errors of different vowels are illustrated in Fig. 7 for the normal subjects and the laryngectomee. The results showed significant differences in the control errors across vowels. Especially for the normal subjects, the vowel /ɑ/ had the highest control error (98.3%) at the voice initiation, while the vowel /u/ got the lowest control error (34.2%). And it needed the longest time for the vowel /ɑ/ (790 ms) to reduce the error below 5%, while the shortest time for the vowel /u/ (165 ms). These outcomes might be explained by the different changing progresses due to the different lip shapes of vowels. Comparing with the silence state, the lip shape difference of vowel /ɑ/ was larger than that of vowel /u/, thus, the progress of lip changing was longer for the vowel /ɑ/ than vowel /u/. By contrast, the vowel differences were smaller in the case of the laryngectomee, because the earlier beginning of changing progress resulted in a smaller lip shape difference from the target vowel and lower control errors at the voice initiation.

For the termination period, the control errors of all participants were approximate to zero. This outcome might be explained by the slower lips closing than voice termination for both normal subjects and laryngectomee (Wu et al., 2013), which guaranteed a stable lip shape and low control error before the termination of vowel phonation. Therefore, all participants achieved good control performance of SGVS-EL in the duration period and termination period, and laryngectomee acquired better control in initiation period than normal subjects.

4.2.2. *Influence of the consonant*

In daily communication, the consonant is a necessary phoneme to express a meaning word, and may have an impact on the vowel control. Fig. 8 shows the real time vowel identification and voice source control of SGVS-EL in the case of CVC words. It was obvious that the voice source could be correctly controlled during phonation of a

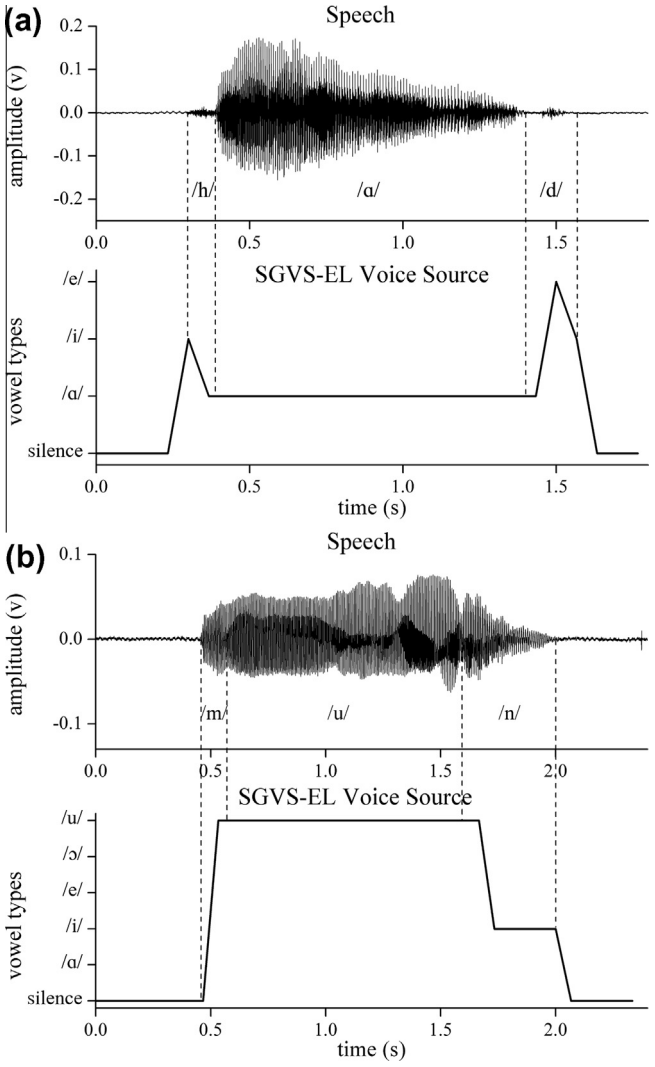

Fig. 8. Real time identification and voice source control of SGVS-EL during phonation of CVC words. (a) The word/hɑ:d/. (b) The word /mu:n/. The section of each phoneme is marked in the graphs.

CVC word, especially in the vowel duration period. However, the control errors also happened at the transition period from consonant to vowel or from vowel to consonant. And the lip shape differences between the consonant and silence state might result in a different control performance from the case of single vowel production.

To study the influence of the consonants, the vowel control errors of the CVC words were measured and compared with that of single vowel. The influences were summarized into three categories: '+' standing for significantly higher control errors than single vowel, '−' standing for significantly lower control errors, and '±' standing for equal control errors. For word-terminal consonants, because the lip changing was always later than the vowel termination, the control errors were still close to zeros, the same as the results in single vowel phonation. Thus, only the influence of word-initial consonants is listed in Table 2.

The results showed no significant difference between the normal subjects and laryngectomee, and decreased vowel control errors for most word-initial consonants. This outcome indicated that the word-initial consonant could improve the performance of vowel control, which might be closely related with the lip shape of the consonant. Because most consonants were produced with open-mouth, which might result in a smaller difference from the target vowel at the vowel initiation. To the contrary, the consonants with close-mouth (/p/, /b/, /m/, /f/) were easily identified to silence state, which made the same situation as the single vowel.

Consequently, all participants can perform a good control of vowel-specific sources for word production, providing a great potential for applying SGVS-EL system in continuous speaking.

4.3. Intelligibility evaluation

4.3.1. Vowel intelligibility

Given that the vowel-specific voice sources of SGVS-EL was intended to improve the vowel quality, the vowel intelligibility was an important aspect for the subjective evaluation. The results of vowel intelligibility are summarized in Table 3. The entries in the table are the mean percentages of correct responses to vowels in words produced with normal voice, SGVS-EL and commercial EL. The numbers in parentheses are standard deviations of the means. As the reference value, the overall vowel intelligibility of normal speech was 98.45%.

The overall vowel intelligibility of SGVS-EL was 90.83% for the normal subjects and 89.76% for the laryngectomee, which was significantly higher than that of commercial EL (60.77% for the normal subjects and 56.19% for the laryngectomee). This result indicated that the vowels produced by SGVS-EL was more intelligible than the commercial EL vowels, which might be closely related with the improved acoustic characteristics of vowels produced with supra-glottal voice source, including the enhanced low-frequency energy, corrected formants, and compensated spectral zeros (Wu et al., 2013). Especially, the first two formants played an important role in vowel perception (Carlson et al., 1975; Traunmuller and Lacerda, 1987). Fig. 9 presents an view of formant frequency by plotting the average F1 and F2 values of five vowels produced by

Table 2
The influence of word-initial consonants on vowel control.

Subject	/l/	/m/	/n/	/f/	/s/	/ʃ/	/h/	/p/	/b/	/t/	/d/	/k/	/g/	/ʧ/
Normal	−	±	−	±	−	−	−	±	±	−	−	−	−	−
Laryngectomee	−	±	−	±	−	−	−	±	±	−	−	−	−	−

The sign '+' means a higher control error than single vowel; the sign '−' means a lower control error; the sign '±' means an equal control error.

Table 3
Intelligibility scores (%) of vowels produced with normal voice, SGVS-EL and commercial EL.

Vowel	Normal speech	Normal subjects		Laryngectomee	
		SGVS-EL	Commercial EL	SGVS-EL	Commercial EL
/ɑ/	98.21 (1.90)	86.90 (5.34)	60.12 (4.93)	89.58 (4.59)	58.33 (6.20)
/i/	99.40 (1.01)	92.86 (4.83)	52.38 (5.54)	92.86 (4.94)	57.74 (6.40)
/e/	99.11 (1.21)	93.45 (5.26)	69.05 (5.34)	92.26 (6.13)	57.44 (5.10)
/ɔ/	96.73 (2.97)	88.10 (4.43)	61.61 (6.48)	89.88 (5.15)	58.63 (4.47)
/u/	98.81 (1.36)	92.86 (5.67)	60.71 (6.14)	84.23 (5.16)	48.81 (6.11)
Means	98.45	90.83	60.77	89.76	56.19

The numbers in parentheses are standard deviations of the means.

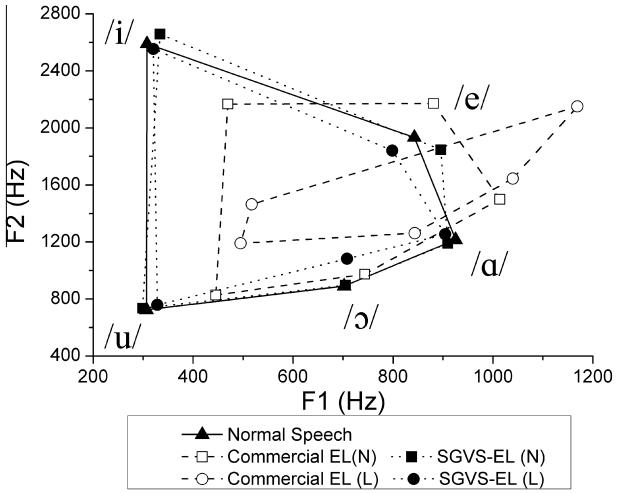

Fig. 9. F1 vs. F2 of five vowels for different subjects and different ELs. Each point represents the average frequencies of F1 and F2 of the EL speech under specific condition.

SGVS-EL and commercial EL for different subjects. It was obvious that the formants of the SGVS-EL speech (solid square and circle) were close to that of the normal speech (solid triangle), while the commercial EL speech formants (hollow square and circle) deviated far from the normal speech. Therefore, using the SGVS-EL system, all participants could produce a vowel with improved acoustic properties and high intelligibility.

The results in Table 3 showed that the vowel intelligibility of SGVS-EL was still lower than that of the normal speech, which might be due to the control errors occurred in the initiation period. However, the impact of the control errors upon the vowel intelligibility was not significant from the following aspects. First, the vowel intelligibility of SGVS-EL was not significantly different across vowels (*ANOVA*, $P > 0.5$). This outcome indicated that the vowel intelligibility was not appreciably influenced by the vowel differences of the control errors. Second, the vowel intelligibility of SGVS-EL was not significantly different across subjects (*T-test*, $P > 0.5$). This outcome indicated that the vowel intelligibility was not appreciably influenced by the individual differences of the control errors. Consequently, the control performance of SGVS-EL was acceptable to reconstruct an intelligible vowel.

4.3.2. Word intelligibility

To comprehensively evaluate the quality of SGVS-EL speech, the consonant and word intelligibility were calculated based on the mean percentages of correct responses to consonants and words for all listeners. The results are illustrated in Fig. 10.

Graph (a) shows the consonant intelligibility scores of SGVS-EL and commercial EL for different subjects. For

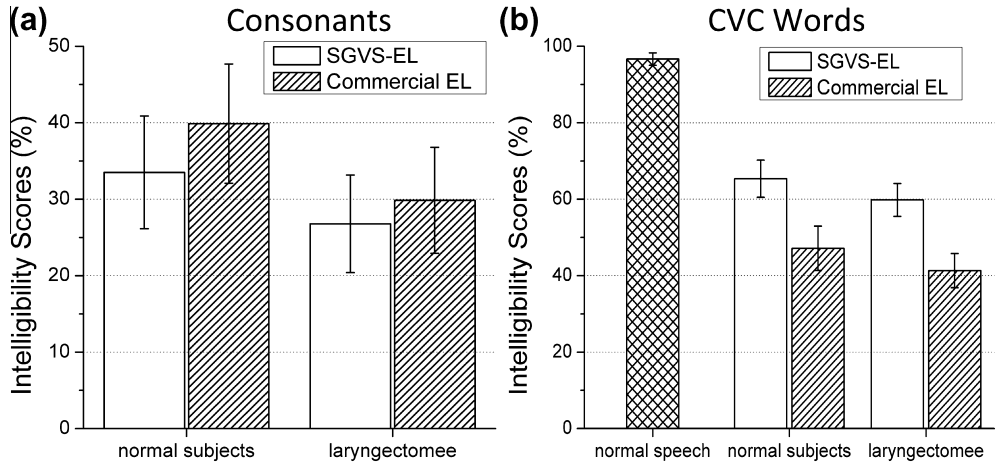

Fig. 10. Intelligibility scores of consonants (a) and CVC words (b) produced with normal voice, SGVS-EL, and commercial EL for different subjects.

SGVS-EL, the average consonant intelligibility score was 33.5% for the normal subjects and 26.8% for the laryngectomee, which was lower than that of commercial EL (39.9% for the normal subjects and 29.9% for the laryngectomee). This outcome indicated that subject using SGVS-EL produced a less intelligible consonant than commercial EL, which might be related with the on/off control performance of SGVS-EL. Previous work demonstrated that the on/off control of ANN method reduced the intelligibility of both word-initial and word-terminal consonants (Wu et al., 2013).

Graph (b) shows the word intelligibility scores of normal voice, SGVS-EL, and commercial EL for different subjects. As a reference value, the word intelligibility score of normal speech was 96.6%. For SGVS-EL, the word intelligibility scores of the normal subjects and laryngectomee were 65.4% and 59.8%, which was significantly higher than that of commercial EL (47.1% for the normal subjects and 41.3% for laryngectomee, *T-test*, $P < 0.05$). Although the consonant of SGVS-EL was less intelligible than commercial EL, all participant using SGVS-EL could still produce a more intelligible word than commercial EL. Therefore, SGVS-EL with controlled vowel-specific voice sources was feasible to improve the intelligibility of EL speech.

In addition, the phonation training might influence the coarticulatory effect and consonant phonation while stabilizing their lips shapes according to the vowels. However, the high intelligibility of normal speech indicates that the training has no significant effect on intelligibility of natural speech. Furthermore, the improper consonant voice source is the main reason for the low consonant intelligibility and unnatural coarticulatory characteristics (Weiss et al., 1979). Thus, the training effect on the coarticulation and consonant intelligibility is less significant in current work.

5. Conclusion

In this paper, an experimental electrolarynx (SGVS-EL) system was designed to implement a real time vowel identification based on visual lips information to control a vowel-specific voice source. Through the training protocol, the artificial neural network (ANN) was able to acquire high accuracy of vowel identification for all participants. Then, the assessment of SGVS-EL voicing control showed an acceptable control performance of vowel-specific voice sources in real time. Especially, the laryngectomee had less control errors than the normal subjects because of the individual phonation habit. Furthermore, the word-initial consonant had a positive influence on the vowel identification to reduce the control errors of CVC words production. Finally, all subjects using SGVS-EL produced more intelligible vowels and words than commercial EL. The results demonstrated that the control method was feasible to regulate the vowel-specific voice source for real time communication, and the SGVS-EL was effective to reconstruct an EL speech with improved quality and high intelligibility. Finally, the shortage of this work is only one laryngectomee in the test group, because low number of laryngectomees are willing to participate in an effortful and time-consuming study. At the moment, the low number of participants is still adequate to demonstrate the method. Following works with mobile device will focus on its applicability on a bigger laryngectomees' cohort in daily life.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 11274250 and 61271087, and Research Fund for the Doctoral Program of Higher Education of China under Grand 20120201110049. The authors would like to express special appreciation to Ye Wenyuan and Rui Baozhen for their works in the experiments.

References

Pawar, P.V., Sayed, S.I., Kazi, R., Jagade, M.V., 2008. Current status and future prospects in prosthetic voice rehabilitation following laryngectomy. J. Can. Res. Ther. 4, 186–191.

Carr, M.M., Schmidbauer, J.A., Majaess, L., Smith, R.L., 2000. Communication after laryngectomy: an assessment of quality of life. Otolaryngol. Head Neck Surg. 122, 39–43.

Clements, K.S., Rassekh, C.H., Seikaly, H., Hokanson, J.A., Calhoun, K.H., 1997. Communication after laryngectomy: an assessment of patient satisfaction. Arch. Otolaryngol. Head Neck Surg. 123, 493–496.

Morris, H.L., Smith, A.E., Van Demark, D.R., Maves, M.D., 1992. Communication status following laryngectomy: the Iowa experience 1984–1987. Ann. Otol. Rhinol. Laryngol. 101, 503–510.

Hillman, R.E., Walsh, M.J., Wolf, G.T., Fisher, S.G., Hong, W.K., 1998. Functional outcomes following treatment for advanced laryngeal cancer. Part I-Voice preservation in advanced laryngeal cancer. Part II-Laryngectomy rehabilitation: the state of the art in the VA System. Ann. Otol. Rhinol. Laryngol. 172, 1–27, Suppl.

Meltzner, G.S., Hillman, R.E., Heaton, J.T., Houston, K.M., Kobler, J.B., 2005. Electrolaryngeal speech: The state of the art and future directions for development. Contemporary considerations in the treatment and rehabilitation of head and neck cancer: voice, speech, and swallowing., 571–590.

Weiss, M.S., Yeni-Komshian, G.H., Heinz, J.M., 1979. Acoustical and perceptual characteristics of speech produced with an electronic artificial larynx. J. Acoust. Soc. Am. 65, 1298–1308.

Qi, Y., Weinberg, B., 1991. Low-frequency energy deficit in electrolaryngeal speech. J. Speech Hear. Res. 34, 1250–1256.

Meltzner, G.S., 2003. Perceptual and acoustic impacts of aberrant properties of electrolaryngeal speech. Doctoral dissertation, Massachusetts Institute of Technology.

Wu, L., Wan, C.Y., Wang, S.P., Wan, M.X., 2013. Improvement of electrolaryngeal speech quality using a supra-glottal voice source with compensation of vocal tract characteristics. IEEE Trans. Biomed. Eng. 60, 1965–1974.

Uemi, N., Ifukube, T., Takahashi, M., Matsushima, J., 1995. Development of an electrolarynx having a pitch frequency control function by using expiration pressure. Jpn. Soc. Med. Biol. Eng. 33, 7–14.

Takahashi, H., Nakao, M., Kikuchi, Y., Kaga, K., 2008. Intra-oral pressure-based voicing control of electrolaryngeal speech with intra-oral vibrator. J. Voice 22, 420–429.

Goldstein, E.A., Heaton, J.T., Kobler, J.B., Stanley, G.B., Hillman, R.E., 2004. Design and implementation of a hands-free electrolarynx device controlled by neck strap muscle electromyographic activity. IEEE Trans. Biomed. Eng. 51, 325–332.

Goldstein, E.A., Heaton, J.T., Stepp, C.E., Hillman, R.E., 2007. Training effects on speech production using a hands-free electromyographically controlled electrolarynx. J. Speech Hear. Res. 50, 335–351.

Stepp, C.E., Heaton, J.T., Rolland, R.G., Hillman, R.E., 2009. Neck and face surface electromyography for prosthetic voice control after total laryngectomy. IEEE Trans. Neural Syst. Rehabil. Eng. 17, 146–155.

Dupon, S., Luettin, J., 2000. Audio-visual speech modeling for continuous speech recognition. IEEE Trans. Multimedia 2, 141–151.

Neti, C., Potaminanos, G., Luettin, J., Matthews, I., Glotin, H., Vergyri, D., Sison, J., Mashari, A., Zhou, J., 2000. Audio-visual speech recognition, Final Workshop 2000 Report, 764.

Massaro, D.W., Cohen, M.M., 1983. Evaluation and integration of visual and auditory information in speech perception. J. Exp. Psychol. Human 9, 753–771.

Evitts, P.M., Portugal, L., Van Dine, A., Holler, A., 2010. Effects of audio-visual information on the intelligibility of alaryngeal speech. J. Commun. Disord. 43, 92–104.

Summerfield, Q., 1992. Lipreading and audio-visual speech perception. Philos. Trans. R. Soc. Lond. B 335, 71–78.

Wan, C.Y., Wu, L., Wu, H.X., Wang, S.P., Wan, M.X., 2012. Assessment of a method for the automatic on/off control of an electrolarynx via lip deformation. J. Voice 26, pp. 674e21–30.

Wu, L., Wan, C.Y., Wang, S.P., Wan, M.X., 2013. Development and evaluation of on/off control for electrolaryngeal speech via artificial neural network based on visual information of lips. J. Voice 27, pp. 259e7–16.

Umeda, N., 1977. Consonant duration in American English. J. Acoust. Soc. Am. 61, 846–858.

Cho, T., Ladefoged, P., 1999. Variations and universals in VOT: evidence from 18 languages. J. Phonetics 95, 207–229.

Carlson, R., Fant, G., Granstrom, B., 1975. Two-formant models, pitch and vowel perception. Auditory Anal. Percept. Speech, 55–82.

Traunmuller, H., Lacerda, F., 1987. Perceptual relativity in identification of two-formant vowels. Speech Commun. 6, 143–157.