Improvement of Electrolaryngeal Speech Quality Using a Supraglottal Voice Source With Compensation of Vocal Tract Characteristics

Liang Wu, Congying Wan, Supin Wang, and Mingxi Wan*

***Abstract*—Electrolarynx (EL) is a medical speech-recovery device designed for patients who have lost their original voice box due to laryngeal cancer. As a substitute for human larynx, the current commercial EL voice source cannot reconstruct natural EL speech under laryngectomy conditions. To eliminate the abnormal acoustic properties of EL speech, a supraglottal voice source with compensation of vocal tract characteristics was proposed and provided through an experimental EL(SGVS-EL) system. The acoustic analyses of simulated EL speech and reconstructed EL speech produced with different voice sources were performed in the normal subject and laryngectomee. The results indicated that the supraglottal voice source was successful in improving the acoustic properties of EL speech by enhancing low-frequency energy, correcting the shifted formants to normal range, and eliminating the visible spectral zeros. Both normal subject and laryngectomee also produced more natural vowels using SGVS-EL than commercial EL, even if the vocal tract parameter was substituted and the supraglottal voice source was biased to a certain degree. Therefore, supraglottal voice source is a feasible and effective approach to improving the acoustic quality of EL speech.**

***Index Terms*—Acoustic analysis, electrolarynx, speech quality, supraglottal voice source.**

I. Introduction

THOUSANDS of people have to undergo total laryngectomy as a treatment of laryngeal cancer, with the removal of the whole larynx and surrounding tissue, therefore resulting in the loss of the voice source and partial natural pathway for speech production. However, given that most of the vocal tract is still reserved, a laryngeal speech can be produced by the electrolarynx (EL). The EL is a handheld, battery-powered electromechanical transducer, which provides a mechanical sound as a substitution source for voice rehabilitation. Although the EL has been widely used for daily communication due to the advantages of easy learning and little maintenance, the unnatural quality of EL speech is a serious defect, which reduces its intelligibility and acceptability [1], [2].

A number of studies have revealed the abnormal acoustic properties of the EL speech. Weiss *et al.* indicated the presence of a significant low-frequency energy deficit below 500 Hz in the EL speech compared with normal speech [3]. Qi and Weinberg demonstrated that the lack of low-frequency energy in the commercial voice source contributed to the unnatural quality of EL speech and reduced its intelligibility [4]. Meltzner compared the laryngectomee's EL speech with his/her normal speech before the surgery, and reported that postlaryngectomy speech had unnatural spectra from the corresponding prelaryngectomy one, such as higher formant frequencies, narrower formant bandwidths, reduced low-frequency energy, and different relative formant amplitudes [5]. Moreover, Meltzner [5], and Myrick and Yantorno [6] found visible spectral zeros in the EL speech.

All the differences between EL speech and normal speech are due to the vocal tract system changes caused by the surgery and EL placement. For pathological anatomy, the laryngectomy not only removes the larynx, but also resects part of the supraglottal vocal tract. This procedure indicates that the reserved vocal tract is shorter, and the vocal tract resonance is changed. For EL use, the neck-type EL is placed against the skin on one side of the neck under the chin, and the vibration is conducted into the laryngopharynx instead of the end of the reserved vocal tract; that is, the lower part of the vocal tract acts as back cavity and introduces spectral zeros into the vocal tract transfer function [6].

There have been some approaches reported to eliminating the abnormal acoustic properties in the EL speech. Qi and Weinberg developed an optimal second-order low-pass filter to compensate for the low frequency deficit [4]. Myrick and Yantorno designed an all pole inverse filter based on the zeros measured in the spectrum of EL speech to compensate for the effects of the zeros [6]. Ma *et al.* used cepstral analysis of speech to replace the EL excitation signal with a normal speech excitation signal while keeping the vocal tract information constant [7]. All the reports demonstrated potential improvements in EL speech quality; however, these postprocessing methods were not the fundamental solutions to the problem and could not fulfill the real-time conversation.

The EL primarily provides an alternative voice source, which is vital for regaining natural speech. However, generating a

Manuscript received November 6, 2012; revised November 23, 2012 and February 6, 2013; accepted February 7, 2013. Date of publication February 12, 2013; date of current version June 24, 2013. This work was supported in part by the National Natural Science Foundation of China under Grant 11274250 and Grant 61271087, and in part by the Research Fund for the Doctoral Program of Higher Education of China under Grant 20120201110049. *Asterisk indicates corresponding author.*

L. Wu, C. Wan, and S. Wang are with the Key Laboratory of Biomedical Information Engineering of Ministry of Education and Department of Biomedical Engineering, School of Life Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: liangliang26@gmail.com; wancy128@stu.xjtu.edu.cn; spwang@mail.xjtu.edu.cn).

*M. Wan is with the Key Laboratory of Biomedical Information Engineering of Ministry of Education and Department of Biomedical Engineering, School of Life Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: mxwan@mail.xjtu.edu.cn).

Digital Object Identifier 10.1109/TBME.2013.2246789

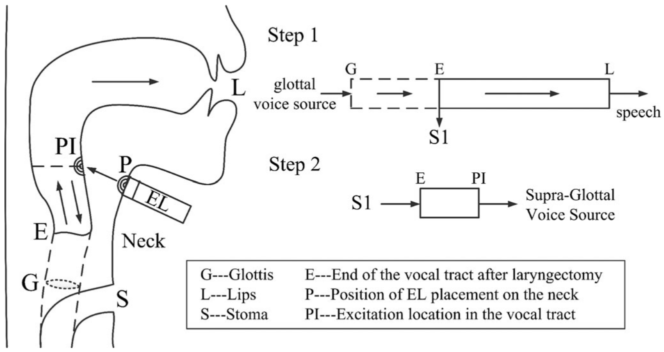
Step 1
G
E
L
glottal voice source
speech
S1
Step 2
E
PI
S1
Supra-Glottal Voice Source
L
PI
P
EL
E
Neck
G
S
G---Glottis
E---End of the vocal tract after laryngectomy
L---Lips
P---Position of EL placement on the neck
S---Stoma
PI---Excitation location in the vocal tract

Fig. 1. Synthetic schematic of the supraglottal voice source. Left image illustrates the sagittal plane of the vocal tract and EL speech production under laryngectomy conditions. The dash line represents the original vocal tract before the laryngectomy. Right image shows the two compensation steps for the synthesis of the supraglottal voice source.

proper voice source to meet the needs of EL speech reconstruction under laryngectomy conditions has not been intensively studied. Mainly two kinds of EL voice sources are currently available according to different EL vibrators. For the conventional nonlinear transducer EL, such as the Bell and Servox EL, the driving signal is a train of square pulses or impulses alternating in sign, which results in a voice source with a train of sharp impulses followed by highly damped oscillations [4]. The EL speech produced with this voice source has the low-frequency energy deficit and abnormal acoustic properties [4], [5]. To provide an arbitrary driving waveform, Houston [8] and Ooe [9] designed linear transducers to provide a glottal voice source. Recent linear transducers might provide a normal glottal voice source without low-frequency deficit, but could not eliminate the other abnormal acoustic properties of EL speech due to the changes in vocal tract structure.

Current EL voice sources cannot produce natural EL speech because of the lack of consideration for the removed vocal tract and back cavity. To eliminate abnormal acoustic properties in the EL speech, we developed a supraglottal voice source with compensation of vocal tract characteristics to adapt the requirements of EL speech production. The supraglottal voice source was generated and provided through an experimental EL (SGVS-EL) system. Then, the EL speech produced with the supraglottal voice source was simulated and reconstructed in the cases of the normal subject and laryngectomee. Finally, to evaluate the contribution of the supraglottal voice source to the improvement of EL speech quality, the acoustic properties of EL speech produced with different voice sources were analyzed and compared with those of the normal speech.

II. Methods

A. Synthesis of Supraglottal Voice Source

As shown in the left image of Fig. 1, supraglottal voice source (at point PI) was an improved glottal voice source (at point G) with compensation of the removed vocal tract (from G to E) and back cavity (from E to PI). Considering that the compensation was closely related with physiological structure and speech production process, speech synthesis technique was introduced to synthesize a supraglottal voice source based on the glottal model and vocal tract model.

1) Glottal Model: Liljencrants–Fant (LF) Model: The LF-model developed by Fant *et al.* [10] was used for the glottal voice source in our method, because of its flexibility in matching all common phonations and convenience in determining the parameters for digital implementations. The LF-model is a time-domain model composed of two sections with the following forms:

$$\begin{cases} u'_g(t) = E_0 e^{\alpha t} \sin(\omega_g t) & (0 \le t \le t_e) \\ u'_g(t) = -\left(\dfrac{E_e}{\varepsilon t_a}\right)\left[e^{-\varepsilon(t-t_e)} - e^{-\varepsilon(t_c-t_e)}\right] & (t_e \le t \le t_c). \end{cases} \tag{1}$$

The LF-model is specified by five parameters, including one amplitude parameter E_e of the negative peak and four timing parameters $\{t_p, t_e, t_a, t_c\}$, which represent the instant of maximum glottal flow, the instant of the negative peak, the time constant of the exponential curve, and the pitch period, respectively. All the other direct synthesis parameters $\{E_0, \alpha, \omega_g, \varepsilon\}$ in (1) can be derived from the timing parameters through the method discussed by Fant *et al.* [10].

2) Vocal Tract Model: 1-D Digital Waveguide Model: Based on the assumptions of straightening the vocal tract and planar wave motion, a one-dimensional (1-D) digital waveguide model [11] was used to simulate wave propagation inside the vocal tract for the compensation.

In the waveguide model, the tract is sampled and represented as a series of connected cylindrical tubes with different sectional areas. Inside the tube section i, according to the d'Alembert solution, the 1-D wave motion is composed of left-going u_i^- and right-going u_i^+ velocity components, and the total velocity is the sum of the two components. Meanwhile, at the junction between tube sections i and $i+1$, the Kelly–Lochabum (KL) junction is used to scatter approaching signals according to the difference in impedance Z, which is the reciprocal of cross-sectional area A. Thus, the scattering equations from a KL junction and a reflection coefficient r_i are defined in (2).

$$\begin{cases} u_{i+1}^+ = (1-r_i)u_i^+ - r_i u_{i+1}^- = u_i^+ - r_i(u_i^+ + u_{i+1}^-) \\ u_i^- = (1+r_i)u_{i+1}^- + r_i u_i^+ = u_{i+1}^- + r_i(u_i^+ + u_{i+1}^-) \end{cases}$$
$$r_i = \frac{A_i - A_{i+1}}{A_i + A_{i+1}}. \tag{2}$$

Specially, at the tract terminals, the glottal end is modeled as closed and lip end as opened, with the equations in

$$\begin{cases} \text{glottis}: \; u_1^+ = \dfrac{1-r_g}{2}u_g - r_g u_1^- \\ \qquad\qquad = \dfrac{1}{2}u_g - r_g\left(\dfrac{1}{2}u_g + u_1^-\right) \qquad r_g \approx -1 \\ \text{lips}: \quad u_{out} = (1-r_N)u_N^+ = u_N^+ - u_N^- \quad r_N \approx -1. \end{cases} \tag{3}$$

The u_g represents the input signal at the glottal end, and u_{out} represents the output signal at the lip end. The r_g and r_N are the reflection coefficients at the glottal and lip ends, respectively.

3) Synthetic Procedures: Based on the LF-model and waveguide model, the synthesis of a supraglottal voice source

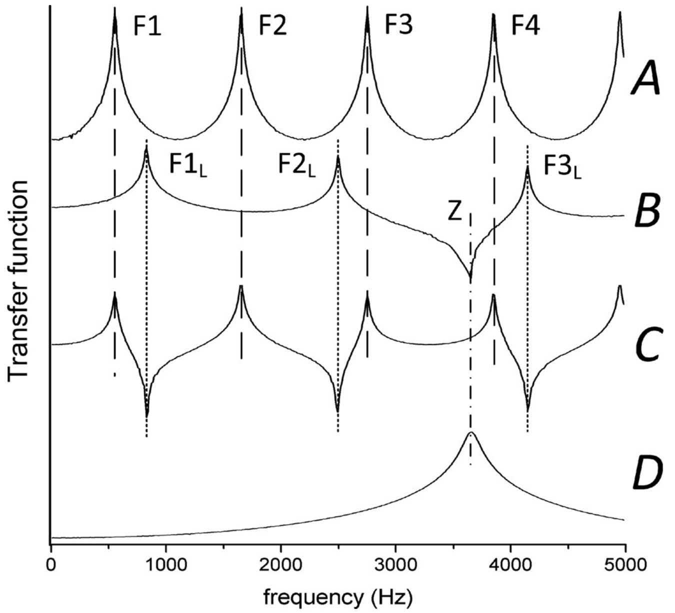

Fig. 2. Acoustic rationale of the vocal tract compensation. (a) Transfer function of the original vocal tract before the laryngectomy. The *F1*, *F2*, *F3*, and *F4* represent the first four formants of the normal speech. (b) Transfer function of the reserved vocal tract after the laryngectomy. The $F1_{L}$, $F2_{L}$, $F3_{L}$, and Z represent the abnormal formants and the spectral zero in the EL speech. (c) Transfer function of the compensated vocal tract for step 1. (d) Transfer function of the back cavity for step 2.

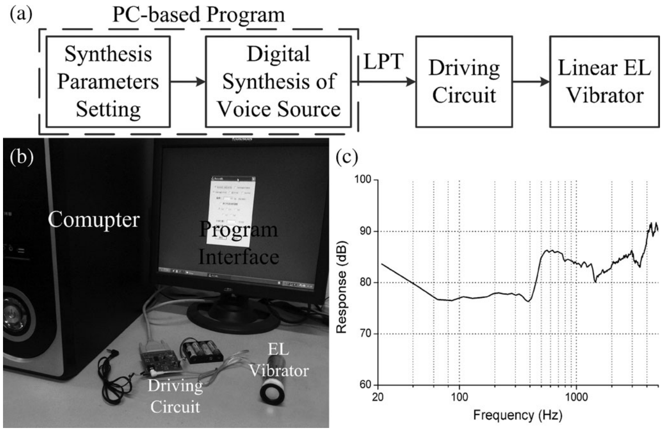

Fig. 3. Experimental electrolarynx (SGVS-EL) system. (a) Schematic diagram of the system. (b) Photo of the system components. (c) Frequency response curve of the linear vibrator with a sweep frequency speed of 100 mm/s.

consisted of two compensation steps as shown in the right image of Fig. 1.

Step 1: Compensate the characteristics of the removed vocal tract (from G to E). The whole original vocal tract (from G to L) was modeled with closed glottal end and opened lip end. The glottal voice source of the LF-model was inputted into the vocal tract at glottis (point G), and the wave propagated to the lips (point L) inside the intact vocal tract. Then, signal S1 at E point was extracted as the output signal of this step. This step was unnecessary for the normal subject speaking with the EL.

Step 2: Compensate the spectra zeros of the back cavity. For the laryngectomee, the back cavity was modeled from point E (closed end) to point PI (opened end). Signal S1 was inputted into the back cavity at point E, and the output signal at point PI was the supraglottal voice source. For the normal subject, the back cavity was modeled from point G (closed end) to point PI (opened end), and the input signal was the glottal voice source.

Assuming that the vocal tract is a uniform tube, the transfer functions of different vocal tracts are shown in Fig. 2 to illustrate the acoustic rationale of each compensation. For step 1, curve C was the transfer function of the compensated vocal tract. The zero-point frequencies of curve C were equal to the formant frequencies ($F1_{L}$, $F2_{L}$, and $F3_{L}$) of the reserved vocal tract (curve B). This condition will eliminate the abnormal characteristics of the reserved vocal tract. Moreover, the peak frequencies of curve C were equal to the formant frequencies (*F1, F2, F3,* and *F4*) of the original vocal tract (curve A). This condition will reconstruct the normal acoustic characteristics of the original vocal tract. Thus, this step can correct the shifted formants resulting from the shortening of the vocal tract that is caused by the laryngectomy surgery. For step 2, curve D showed the transfer function of the compensated back cavity. The peak frequency of curve D was equal to the zero-point frequency (Z) of the reserved vocal tract (curve B), indicating that the influence of spectral zeros can be eliminated by this step. Therefore, the supraglottal voice source has the potential to improve the acoustic quality of EL speech from the acoustic principle perspective.

The parameters of vocal tract were determined according to the laryngectomy surgery and EL placement. The details of the parameter extraction are described in the experimental section.

B. Experimental Electrolarynx (SGVS-EL) System

The schematic diagram and components of the experimental EL (SGVS-EL) system are shown in Fig. 3(a) and (b). The first part is a PC-based (Lenovo, China) program developed for the digital synthesis of the EL voice source in real time. Users can choose the voice source type and synthesis parameters through an interactive interface. The second part is a driving circuit and a wearable linear EL vibrator. The driving circuit receives the digital signals of the voice source via a parallel port from the computer, and then performs D–A conversion and power amplification to drive the linear vibrator. The EL vibrator is made of a mini-speaker (Somic SN-401, China) with a 3-cm diameter.

The frequency response curve of the EL vibrator is shown in Fig. 3(c). Although the response in the low-frequency region (especially from 60 to 300 Hz) was about 7 dB lower than that in the high-frequency region (from 300 to 5000 Hz), the response curve was adequately flat in a wide interested region (from 20 to 5000 Hz). Given that the neck tissue transfer function was characterized as a low-pass filter by Meltzner *et al.* [12], the linear EL vibrator might realize an approximate compensation of the sound transmission characteristics of the neck. Moreover, the intensity level of speech produced by SGVS-EL was 60 ± 3 dB SPL.

III. Experiments

A. Simulation Experiment

The simulation of EL speech reconstruction was implemented using MATLAB program based on source-filter theory. The voice source was inputted into the vocal tract at the excitation location (see point PI in Fig. 1). The waveguide model was used to simulate the wave propagation in the vocal tract.

Three types of voice sources, namely, supraglottal voice source, commercial EL voice source (Servox digital, Servona GmbH), and glottal voice source (LF-model), were selected to simulate EL speech production in the normal subject and laryngectomized subject.

Two typical vowels were selected: high vowel /i/ as in "beat", and low vowel /ɔ/ as in "bought", because their vocal tract shapes and acoustic properties were quite different from each other. The vocal tract was shaped according to the area functions from the magnetic resonance imaging reported by Story *et al.* [13]. For the normal subject, the glottal end of the vocal tract was set as closed. In the case of the laryngectomee, the vocal tract was truncated at the lower edge of the pharyngeal region (see point E in Fig. 1, 3.97 cm from glottis). The excitation location was set at the middle of the pharyngeal region (see point PI in Fig. 1, 7.14 cm from glottis), which was common among laryngectomy patients [5].

In the ideal case, the supraglottal voice source was unbiased, which referred that the excitation location selected for vocal tract compensation in the supraglottal voice source was completely consistent with the exact excitation location of the EL source for speech production. Hence, the acoustic properties of EL speech simulated with unbiased supraglottal voice source were measured to confirm the feasibility of the supraglottal voice source in theoretically improving the acoustic quality of EL speech.

In the actual EL speech production, the supraglottal voice source could not be completely unbiased, indicating the inconsistent compensation of vocal tract with the exact excitation location of the EL source. To study the influence of the biased supraglottal voice source on EL speech quality, five even-distributed locations within 7.14 ± 1.5 cm from glottis were separately excited to produce EL speech, whereas the supraglottal voice source stayed the same with a fixed compensation of vocal tract (7.14 cm from glottis). Finally, the average acoustic properties of EL speech simulated with biased supraglottal voice sources were calculated and compared with normal speech. This procedure aimed to prove that the supraglottal voice source had the potential to be applied in practical EL speech production.

B. SGVS-EL Speech Reconstruction

One male laryngectomee and one male normal speaker participated in the experiment of an EL speech reconstruction. The 76-year-old laryngectomee underwent total laryngectomy because of laryngeal cancer, and had two years experience in using commercial EL. The 25-year-old normal subject was moderately proficient in using the EL.

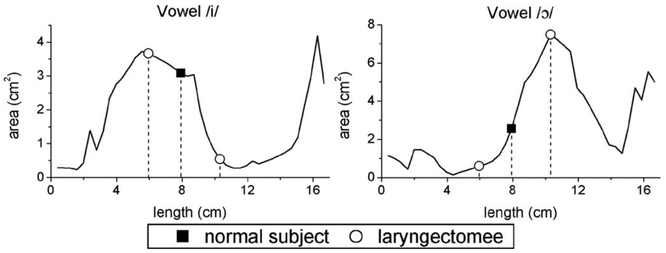

Fig. 4. Extracted vocal tract area functions of the normal subject. The solid square represents the excitation location (see point PI in Fig. 1) in the case of the normal subject. The hollow circles represent the truncated vocal tract (see point E in Fig. 1) and excitation location (see point PI in Fig. 1) in the case of the laryngectomee.

The same sustained vowels, /i/ as in "beat" and /ɔ/ as in "bought", were selected for the experiment. During the actual recording, each subject was instructed to produce each vowel with the normal voice (only for the normal subject), commercial EL voice source (Servox digital, Servona GmbH), and supraglottal voice source, respectively. To eliminate the differences in sound pressure and self-noise between SGVS-EL and commercial EL, the two EL voice sources were both provided by SGVS-EL. In accordance with the condition of the laryngectomee, the normal speaker was requested to hold his glottis closed during speech to eliminate inadvertent voicing and air flow from the lower respiratory tract.

Before this experiment, the vocal tract parameters were estimated based on the acoustic properties of commercial EL speech, and shown in Fig. 4.

1) For the normal subject, the vocal tract area functions were estimated from his normal speech through simulated annealing based on the vocal tract articulatory model [14]. Then, the spectral zeros of commercial EL speech were used to estimate the excitation location (see point PI in Fig. 1) in the extracted vocal tract. In this experiment, the excitation location of the normal subject was approximately 8 cm from glottis in the estimated vocal tract [see Fig. 4], which determined the back cavity for compensation.
2) For most laryngectomized subjects, obtaining their intact vocal tract area functions, which, however, are vital in our method, is impossible. Considering that the vocal tract shapes of different people are similar for the same vowel [13], parameter substitution is reasonable. To implement the supraglottal voice source in laryngectomy conditions, the laryngectomee's vocal tract was substituted by that of the normal subject to synthesize an alternative supraglottal voice source. Thus, the parameters for vocal tract compensation were extracted based on the substituted vocal tract. First, the first formant frequency of commercial EL speech produced by the laryngectomee was extracted to estimate the truncated vocal tract (see point E in Fig. 1), because the first formant frequency was sufficiently low to avoid influence from the spectral zeros of the back cavity. Second, the spectral zeros of commercial EL speech were used to estimate the back cavity and excitation location (see point PI in Fig. 1). In this exper-

iment, points E and PI were approximately 6 and 10 cm, respectively, from glottis in the substituted vocal tract for the laryngectomized subject [see Fig. 4].

All the recordings were performed in a soundproof room. Speech signals were collected using a data acquisition system (BioPac MP150) with a dynamic microphone (Salar M9) mounted 5 cm away from the mouth. The acoustic data were digitized at a 44 100 Hz sampling rate with a 16-bit quantization.

C. Acoustic Analysis

The supraglottal voice source was intended to improve EL speech quality, especially to eliminate the abnormal acoustic properties. Moreover, the acoustic spectrum was an important aspect for the objective evaluation of vowel quality [15]. Thus, in this study, the acoustic properties of the EL speech produced with the supraglottal voice source were mainly analyzed as follows.

A 4096-point Discrete Fourier Transform (DFT) with a 50 ms Hamming window was utilized to calculate the spectrum. Linear predictive (LP) coefficients were calculated for each recording using the autocorrelation method of autoregressive (AR) modeling.

Three aspects of acoustic properties were considered. The first three formant frequencies (*F1, F2,* and *F3*) and amplitudes (*A1, A2,* and *A3*) were measured based on both the DFT spectrum and LP spectrum. The low-frequency energy was evaluated by a normalized quantity (Er*)*. Er was defined as the energy below 500 Hz divided by the energy below 5000 Hz, which was the frequency range of interest. Finally, the frequencies of visible spectral zeros were marked manually.

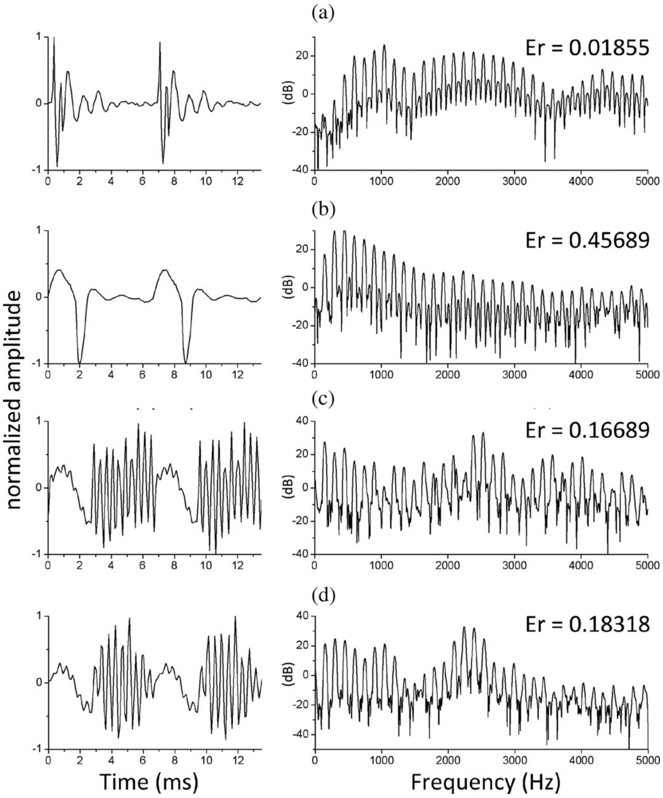

Fig. 5. Normalized acoustic signals (left) and the corresponding spectra (right) of different voice sources provided by SGVS-EL. (a) Commercial EL voice source. (b) Glottal voice source. (c) Supraglottal voice source of vowel /i/ for the laryngectomee. (d) Supraglottal voice source of vowel /ɔ/ for the laryngectomee. The Er value represents the normalized low-frequency energy of each voice source.

IV. Results and Discussion

A. System Output of the Supraglottal Voice Source

Fig. 5 shows the normalized acoustic signals and the corresponding spectra of three different voice sources provided by SGVS-EL. Although the linearity of the transducer was slightly poor in low frequency [see Fig. 3(c)], the distortion of the generated waveform was lower than 3%, which should be acceptable.

The results indicated that the supraglottal voice source displayed significant differences from two other voice sources. In the time domain, the waveform of the supraglottal voice source was more similar to the glottal voice source than the commercial EL voice source, especially in the opening phase of the vibration period, because the supraglottal voice source was developed from the glottal voice source. However, in the closed phase of the vibration period, the waveform of the supraglottal voice source was more fluctuant than the glottal voice source due to the compensation of vocal tract characteristics.

In the frequency domain, the *Er* of the supraglottal voice source (0.16689 for vowel /i/ and 0.18318 for vowel /ɔ/) was much larger than 0.01855 of the commercial EL voice source, which indicated the successful enhancement of low-frequency energy in the supraglottal voice source. Furthermore, the frequency distribution of the supraglottal voice source was different from that of the glottal voice source, which reflected the characteristics of the compensated vocal tract. In particular, the peak frequencies of vowel /i/ were about 300, 2500, and 3500 Hz, which were different from 600, 1100, and 2300 Hz of vowel /ɔ/. This result indicated that the supraglottal voice source was different across vowels because of different vocal tract shapes. Meltzner revealed that the abnormal acoustic properties of EL speech appeared to be vowel dependent [5]. Therefore, vowel-specific supraglottal voice source might be reasonable to reproduce different vowels.

B. Acoustic Analysis of Simulated EL Speech

1) Case of the Unbiased Supraglottal Voice Source: The acoustic properties of EL speech simulated with the unbiased supraglottal voice source are listed in Table I for vowel /i/ and Table II for vowel /ɔ/, comparing with the normal voice, commercial EL voice source, and glottal voice source.

For the normal subject, the data showed unnatural acoustic properties in the EL speech simulated with the commercial EL voice source, but not in that with the supraglottal voice source. First, the $F1$ of the commercial EL voice source was higher than that of the normal speech (by 84.5 Hz for vowel /i/ and 19.3 Hz for vowel /ɔ/), and the $A1$ was lower by 13.8 dB for vowel /i/ and 8.8 dB for vowel /ɔ/. This outcome could be explained by the low-frequency energy deficit in the commercial voice

TABLE I
ACOUSTIC PROPERTIES OF SIMULATED VOWEL /i/ WITH DIFFERENT VOICE SOURCES

Acoustic Properties	Normal Voice	Normal Subject			Laryngectomee		
		Commercial EL Voice Source	Glottal Voice Source	Supra-Glottal Voice Source	Commercial EL Voice Source	Glottal Voice Source	Supra-Glottal Voice Source
$F1$/Hz	271.7	356.2	271	272	387.5	294.9	270.4
$F2$/Hz	2556.9	964.3	--	2554.6	979.4	--	2517.7
$F3$/Hz	3578.3	2513.9	3568	3324.9	3157	3758.3	3559.6
$A1$/dB	35	21.2	37.6	34.2	28.7	40.9	38.2
$A2$/dB	9.7	12.7	--	15.8	17.7	--	14.4
$A3$/dB	24.1	29	9.9	25.1	28.4	4.8	25.1
Er	0.5324	0.0933	0.8466	0.4565	0.2034	0.9064	0.5336
Zeros/Hz	--	2016	--	--	2853	--	--

TABLE II
ACOUSTIC PROPERTIES OF SIMULATED VOWEL /ɔ/ WITH DIFFERENT VOICE SOURCES

Acoustic Properties	Normal Voice	Normal Subject			Laryngectomee		
		Commercial EL Voice Source	Glottal Voice Source	Supra-Glottal Voice Source	Commercial EL Voice Source	Glottal Voice Source	Supra-Glottal Voice Source
$F1$/Hz	612.1	631.4	620.2	614.8	741.7	750.7	619.8
$F2$/Hz	1053.4	1047.8	1057.7	1050.1	1649.5	1604.9	1055.6
$F3$/Hz	2286.7	2227.7	2070.4	2259.5	3158.3	3344.8	2241.5
$A1$/dB	32	23.2	34.6	28.9	28.1	38.7	32.4
$A2$/dB	37.6	35	31.2	39.6	27	14.7	35
$A3$/dB	8.4	13.1	-4.1	9.4	13.8	0.3	6.3
Er	0.2316	0.0184	0.3457	0.2465	0.0351	0.3341	0.2142
Zeros/Hz	--	1139 / 2409	1125 / 2412	--	2398	--	--

source [4], which attenuated the true $F1$ and $A1$ and produced a false $F1$ in higher frequency. In contrast, the Er of the supraglottal voice source (0.4565 for vowel /i/ and 0.2465 for vowel /ɔ/) was higher than that of the commercial EL voice source (0.0933 for vowel /i/ and 0.0184 for vowel /ɔ/). This outcome demonstrated the successful enhancement of low-frequency energy in the EL speech; hence, producing a natural $F1$ and $A1$. Second, a redundant and undesired $F2$ (964.3 Hz) appeared in the simulated vowel /i/ with the commercial EL voice source, which should be related to the abnormal frequency peak at 1 kHz in the spectrum of commercial EL voice source as shown in Fig. 5(a). Thus, the true $F2$ was effectively acting as the $F3$. However, the natural $F2$ was not obviously influenced in the EL speech simulated with the supraglottal voice source, because the frequency peaks of the supraglottal voice source were related to the spectral characteristics of the compensated vocal tract and intended to compensate for the spectral zeros. Consequently, the supraglottal voice source provided with the linear EL vibrator can enhance the low-frequency energy and reconstruct natural formants.

Although the glottal voice source was also capable of enhancing the low-frequency energy and producing natural $F1$ and $A1$, the $F2$ (2556.9 Hz) of normal speech disappeared in the spectrum of simulated vowel /i/ with the glottal voice source. This outcome could be explained by the spectral zeros introduced by placing the voice source away from the end of the vocal tract [5], [6]. Table III shows the theoretical spectral zeros of the back cavity in the simulation experiment. The natural $F2$ (2556.9 Hz) was probably canceled out in simulated vowel /i/ by the spectral zero at 2449 Hz, and the $A3$ of simulated vowel /i/ and /ɔ/ was reduced by the second spectral zeros at 4037 and 2390 Hz, respectively. However, no visible spectral zeros and canceled formants in the EL speech simulated with supraglottal voice source were observed, which demonstrated that the compensation of the back cavity was able to eliminate the effects of spectral zeros. Therefore, the results prove that the supraglottal voice source is feasible to rectify the abnormal acoustic properties of the EL speech produced by the normal subject.

TABLE III
THEORETICAL SPECTRAL ZEROS OF THE BACK CAVITY IN SIMULATION EXPERIMENT

Vowel	Normal Subject		Laryngectomee
/i/	2449Hz	4037Hz	2939Hz
/ɔ/	1195Hz	2390Hz	2455Hz

In the case of the laryngectomized subject, the major difference from the normal subject is the truncation of the vocal tract, which results in a systematic increase in the formant frequencies of the EL speech [4]. In Table II, the obvious observations were that the first three formant frequencies of simulated vowel /ɔ/ with the commercial EL voice source and glottal voice source were higher than that of normal speech (by 129.6, 596.1, and 871.6 Hz for the commercial EL voice source, and 138.6, 551.5, and 1058.1 Hz for the glottal voice source). However, the maximal error of formant frequencies between the supraglottal voice source and normal voice was not more than 50 Hz, which indicated that the characteristic compensation of the removed vocal tract was feasible and effective in eliminating the influence of vocal tract shortening resulting from the laryngectomy surgery.

Furthermore, the abnormal acoustic properties of EL speech simulated with the commercial EL voice source and glottal voice source in the normal subject still existed in the case of laryngectomee, such as the undesired $F2$ (979.4 Hz) and the reduced $A1$ (5.3 dB lower than normal speech) of vowel /i/ with the commercial EL voice source, and the canceled $F2$ (2556.9 Hz) of vowel /i/ with the glottal voice source. Similarly, the supraglottal voice

TABLE IV
AVERAGE ACOUSTIC PROPERTIES OF EL SPEECH SIMULATED WITH BIASED SUPRA-GLOTTAL VOICE SOURCES

Acoustic Properties	Vowel /i/			Vowel /ɔ/		
	Normal Voice	Supra-Glottal Voice Source		Normal Voice	Supra-Glottal Voice Source	
		Normal Subject	Laryngectomee		Normal Subject	Laryngectomee
F1/Hz	271.7	256.68 ± 11.01	270.2 ± 1.20	612.1	613.84 ± 12.45	621.34 ± 9.58
F2/Hz	2556.9	2503.34 ± 32.46	2511.08 ± 8.32	1053.4	1061.74 ± 13.38	950.68 ± 11.05
F3/Hz	3578.3	3417.5 ± 128.85	3543.48 ± 9.50	2286.7	2251.42 ± 27.96	2241 ± 6.16
A1/dB	35	30.6 ± 2.22	37.8 ± 0.93	32	29.06 ± 4.00	32.34 ± 0.40
A2/dB	9.7	14.86 ± 2.35	15.14 ± 1.66	37.6	38.54 ± 3.12	34.7 ± 0.77
A3/dB	24.1	23.32 ± 3.10	26.54 ± 0.97	8.4	8.66 ± 2.44	6.18 ± 1.09
Er	0.5324	0.4654 ± 0.053	0.5034 ± 0.031	0.2316	0.2089 ± 0.022	0.2258 ± 0.019
Zeros/Hz	--	1567 / 2702	--	--	--	--

Five excitation sources were evenly distributed within a 3-cm range around the fixed position of compensation (7.14 -cm from glottis). The entry in the table is the mean and standard deviation of each acoustic property of the EL speech simulated with the five different excitation sources.

source still succeeded in the enhancement of low-frequency energy and compensation of spectral zeros. Therefore, the supraglottal voice source is appropriate and competent to reconstruct a more natural EL speech under laryngectomy conditions.

2) Case of the Biased Supraglottal Voice Source: Five equally spaced excitation locations were selected within a 3-cm range around the fixed position (7.14 cm from glottis) for vocal tract compensation. The average acoustic properties of EL speech simulated with the biased supraglottal voice source are listed in Table IV.

The results showed that the standard deviation of each formant frequency was less than 5%, which indicated that the influence of the biased supraglottal voice source in a certain range was visible but insignificant on the formant frequency. Comparing with normal speech, the average relative errors of first three formant frequencies of the EL speech simulated with the biased supraglottal voice source (2.46% for normal subject and 2.76% for laryngectomee) were less than that of the commercial EL voice source (21.57% for normal subject and 38.66% for laryngectomee) and glottal voice source (14.31% for normal subject and 33.79% for laryngectomee) in Tables I and II. This outcome demonstrated that the biased supraglottal voice source was still able to improve abnormal acoustic properties resulting from the laryngectomy surgery and mid-neck placement of EL.

Most formant amplitude deviations were larger than 10%, which indicated that the influence of the biased supraglottal voice source was greater on the formant amplitude, because the formant amplitude was sensitive to the vocal tract acoustics and the voice source characteristics [5]. Furthermore, the standard deviations of the acoustic properties in the laryngectomee were smaller than in the normal subject by an average of 4.6%, which demonstrated that the influence of the biased supraglottal voice source was smaller in laryngectomy conditions. This result might be related to the shorter vocal tract after the laryngectomy surgery, which weakens the influence of the biased supraglottal voice source.

Visible spectral zeros in simulated vowel /i/ but not in vowel /ɔ/ were observed, which indicated that the influence of the biased supraglottal voice source was different across vowels. This outcome might be explained by the larger variability in the zero frequency of vowel /i/ than vowel /ɔ/ with the excitation source moving because of the larger variability of the cross-sectional areas at the opened end of the back cavity in vowel /i/ [5].

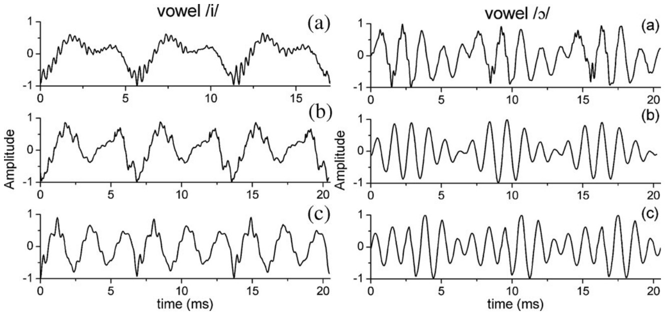

Fig. 6. Reconstructed speech waveforms of vowel /i/ and vowel /ɔ/ in the normal subject. (a) Normal speech. (b) SGVS-EL speech. (c) Commercial EL speech.

In the simulation experiment, the excitation source was assumed as a point source. However, a better model of the EL excitation source should be a distributed source because the EL transducer has a non-zeros size. Nevertheless, the 3-cm range of biased excitation sources was the same as the size of the EL transducer. The average acoustic properties of the EL speech simulated with five equally spaced excitation sources can reflect the general results with multiple adjacent excitation locations. Consequently, the results imply that the supraglottal voice source has the potential to reconstruct a natural EL speech in the practical application.

C. Acoustic Analysis of SGVS-EL Speech

Using the SGVS-EL system, the EL speech was produced with the supraglottal voice source (SGVS-EL speech) and commercial EL voice source (commercial EL speech), respectively. The acoustic waveforms are shown in Fig. 6 for the normal subject and Fig. 7 for the laryngectomized subject. The acoustic properties of SGVS-EL speech are listed in Table V for the normal subject and in Table VI for the laryngectomee, comparing with the commercial EL speech and normal speech.

1) Normal Subject: Fig. 6 shows that the shapes of SGVS-EL speech waveforms were more similar to the normal speech

TABLE V
ACOUSTIC PROPERTIES OF RECONSTRUCTED SPEECH WITH DIFFERENT VOICE SOURCES FOR NORMAL SUBJECT

Acoustic properties	Vowel /i/			Vowel /ɔ/		
	Normal Speech	Commercial EL Speech	SGVS-EL Speech	Normal Speech	Commercial EL Speech	SGVS-EL Speech
F1/Hz	310.2	436.9	294.6	672.8	743.3	704.8
F2/Hz	2555.4	2149	2656.1	915.1	974	920.8
F3/Hz	3285.3	2886.2	3631.3	2949	2856.9	3124.9
A1/dB	33.8	32.1	35.3	34.4	32.1	36.5
A2/dB	16.1	28.7	16.9	28.5	35.9	34.7
A3/dB	18.3	16.9	14.9	20.6	13.5	19.2
Er	0.5062	0.204	0.4916	0.204	0.0495	0.1592
Zeros/Hz	--	1340 - 1467	--	--	3801 - 3828	--

TABLE VI
ACOUSTIC PROPERTIES OF RECONSTRUCTED SPEECH WITH ALTERNATIVE SUPRA-GLOTTAL VOICE SOURCE FOR LARYNGECTOMIZED SUBJECT

Acoustic properties	Vowel /i/			Vowel /ɔ/		
	Normal Speech	Commercial EL Speech	SGVS-EL Speech	Normal Speech	Commercial EL Speech	SGVS-EL Speech
F1/Hz	310.2	517.3	294.6	672.8	783.6	707.8
F2/Hz	2555.4	1169	2525.3	915.1	1198.3	1081.6
F3/Hz	3285.3	2270.7	3468.9	2949	2035.7	2662.2
A1/dB	33.8	28.3	35.7	34.4	29.3	37.7
A2/dB	16.1	21.3	18.3	28.5	22.9	27.1
A3/dB	18.3	18	12.9	20.6	7.9	20.2
Er	0.5062	0.1304	0.426	0.204	0.0585	0.1852
Zeros/Hz	--	1406 - 1661	1187	--	2409 - 3233	--

The column of normal voice represents the acoustic properties of the normal speech produced by the normal subject.

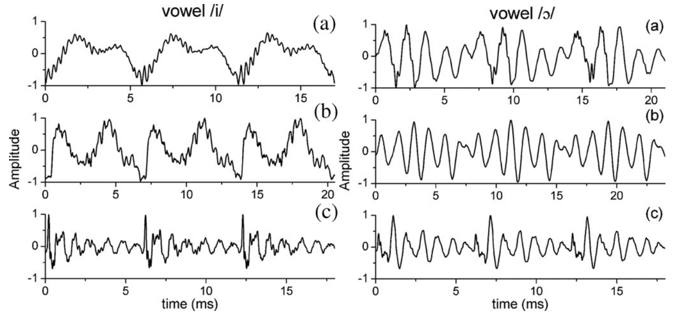

Fig. 7. Reconstructed speech waveforms of vowel /i/ and vowel /ɔ/ in the laryngectomized subject. (a) Normal speech produced by the normal subject. (b) SGVS-EL speech. (c) Commercial EL speech.

than the commercial EL speech, especially for vowel /i/. In Table V, the same as simulation results, the acoustic properties of the commercial EL speech deviated from the normal speech due to the abnormal spectral characteristics of the commercial EL voice source and spectral zeros introduced by the mid-neck placement of the EL.

In contrast, the acoustic quality of SGVS-EL speech was higher than the commercial EL speech in the following two aspects. On the one hand, the errors of the $F1$ and Er between SGVS-EL speech and normal speech were 15.6 Hz, 0.0146 for vowel /i/, and 32 Hz, 0.0448 for vowel /ɔ/, which were smaller than that between the commercial EL speech and normal speech (126.7 Hz, 0.3022 for vowel /i/, and 73.2 Hz, 0.1545 for vowel /ɔ/). This result is mainly due to the better characteristics of the supraglottal voice source without low-frequency energy deficit and the linear output of the EL transducer. On the other hand, the compensation of the back cavity prevents the attenuation of formants from spectral zeros. In the case of vowel /i/, the estimated back cavity should introduce a zero at 2555 Hz in theory, which probably accounts for the $F2$ shifting to 2149 Hz in the commercial EL speech. However, the spectral zeros were well compensated in SGVS-EL speech, which corrected the formants and improved the EL speech quality.

For the commercial EL speech, the results showed that both the waveform and acoustic properties of vowel /ɔ/ were more similar to the normal speech than vowel /i/, which indicated that the commercial EL voice source was merely suitable for the reconstruction of some special vowels because of its own acoustic characteristics. In contrast, both vowels /i/ and /ɔ/ in SGVS-EL speech were as natural as the normal speech, because the supraglottal voice source was vowel dependent. This result demonstrates that the supraglottal voice source is appropriate and feasible for the EL speech reconstruction of any vowel.

However, large errors of the acoustic properties between SGVS-EL speech and normal speech still occurred, especially in high frequency (i.e., $F3$, which was higher by 346 Hz for vowel /i/ and 175.9 Hz for vowel /ɔ/). First, the vocal tract area functions extracted from the normal speech were not entirely accurate, which resulted in a supraglottal voice source with errors. Then, the biased supraglottal voice source that excited the vocal tract at multiple adjacent locations brought new errors into SGVS-EL speech. Nevertheless, the acoustic quality of SGVS-EL speech was still higher than that of the commercial EL speech. The first two formants of SGVS-EL speech were especially closer to the normal speech, which played an important role in vowel perception [5].

2) Laryngectomized Subject: The results in Table VI showed that the acoustic properties of SGVS-EL speech were significantly different from those of the commercial EL speech, which indicated that the alternative supra-glottal voice source had an impact on the reconstructed EL speech. First, the $F1$ (517.3 Hz) of vowel /i/ in the commercial EL speech was higher than the normal range (200 to 400 Hz), and the $F2$ (1169 Hz) was much lower than the normal value (about 2500 Hz) [16]. On the contrary, the $F1$ (294.6 Hz) and $F2$ (2525.3 Hz) of SGVS-EL speech were within the normal ranges, which demonstrated that the compensation of the substituted vocal tract was feasible to correct the shifted formants resulting from the shortening of vocal tract after the surgery. Second, the $A1$ and *Er* of SGVS-EL speech were higher than those of the commercial EL speech (by 7.4 dB and 0.2956 for vowel /i/, 8.4 dB and 0.1267 for vowel /ɔ/), which indicated the effective enhancement of the low-frequency energy. Finally, most of the visible spectral zeros in SGVS-EL speech were eliminated through the compensation of the back cavity. Consequently, the alternative supraglottal voice source can also improve the acoustic properties of the EL speech.

In Fig. 7, the waveforms of SGVS-EL speech were more similar to the normal speech than commercial EL speech. The formant frequencies (especially $F1$ and $F2$) of SGVS-EL speech were also close to those of the normal speech produced by the normal subject, which indicated that the acoustic properties of the reconstructed EL speech could be determined by the vocal tract used for the synthesis of the supraglottal voice source. This condition might be explained by the acoustic rationale of the vocal tract compensation shown in Fig. 2. Step 1 introduced the formants of the substituted vocal tract into the SGVS-EL speech. Thus, controlling the individual features of the EL speech by adjusting the vocal tract parameters might be possible.

In the experiment, considering that the normal subject was 51 years younger than the laryngectomized subject, the substituted vocal tract of the former might be significantly different from that of the latter. Consequently, the estimated parameters for the vocal tract compensation were biased, which synthesized a biased supraglottal voice source and introduced new spectral zeros (i.e., 1187 Hz in vowel /i/) into SGVS-EL speech. These zeros influenced the acoustic quality of the reconstructed EL speech to a certain extent, but did not seriously affect the main formants ($F1$ and $F2$). The reason is that the influence of spectral zeros was weakened by the compensation of the back cavity, and the main formants were strengthened by the compensation of the removed vocal tract. On the other hand, the extracted parameters could not be always realistic. The extracted excitation location in the laryngectomee (10 cm from glottis) was higher than 7 cm of the common position for the EL placement [5]. Nevertheless, the alternative supraglottal voice source could still possibly improve EL speech quality, because the parameters were estimated based on the abnormal properties of the commercial EL speech and represented the characteristics of the reserved vocal tract.

Therefore, the supraglottal voice source using a substituted vocal tract is feasible and effective in eliminating the abnormal acoustic properties and reconstructing natural EL speech under laryngectomy conditions.

D. Applicability of the SGVS-EL System

The compensation of vocal tract characteristics determines that supraglottal voice source is a vowel-specific driving waveform. For daily communication, real-time identification of vowels is essential for using SGVS-EL in continuous speaking. To provide the application possibility of SGVS-EL system in every-day life, we have proposed a real-time method to identify the vowel based on visual information and control the synthesis parameters of supraglottal voice source.

V. Conclusion

To improve the EL speech quality, a supraglottal voice source was synthesized and provided using an SGVS-EL system to reconstruct a natural EL speech. Acoustic properties were measured to evaluate the acoustic quality of the EL speech. The simulation results indicated that the supraglottal voice source could eliminate the abnormal acoustic properties of the EL speech, even if the supraglottal voice source was biased in a small range. First, the acoustic characteristics of the supraglottal voice source enhanced the low-frequency energy of the EL speech. Second, the compensation of the removed vocal tract and back cavity successfully corrected the shifted formants and eliminated the spectral zeros. Then, the reconstruction experiment suggested that the implementation strategy was feasible both in normal and laryngectomy conditions, and successful in improving the acoustic properties of the EL speech. Furthermore, the alternative supraglottal voice source synthesized with the substituted vocal tract indicated a possible way to control the individual features of the EL speech. Therefore, supraglottal voice source is a feasible and effective approach to improving the acoustic quality of EL speech. Our further work will focus on the synthesis control of supraglottal voice source and perceptual evaluation of SGVS-EL speech.

Acknowledgment

The authors would like to thank Y. Wenyuan and R. Baozhen for their cooperation in the experiments.

References

[1] R. E. Hillman, M. J. Walsh, G. T. Wolf, S. G. Fisher, and W. K. Hong, “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. Research speech-language pathologists. Department of Veterans Affairs Laryngeal Cancer Study Group,” *Ann. Otol. Rhinol. Laryngol. Suppl.*, vol. 172, pp. 1–27, 1998.

[2] G. S. Meltzner, R. E. Hillman, J. T. Heaton, K. M. Houston, J. B. Kobler, and Y. Qi, “Electrolaryngeal speech: The state of the art and future directions for development,” in *Contemporary Considerations in theTreatment and Rehabilitation of Head and Neck Cancer: Voice, Speech, and Swallowing*. New York, NY, USA: Barnes & Noble, 2005, pp. 571–590.

[3] M. S. Weiss, G. H. Yeni-Komshian, and J. M. Heinz, “Acoustical and perceptual characteristics of speech produced with an electronic artificial larynx,” *J. Acoust. Soc. Amer.*, vol. 65, pp. 1298–1308, 1979.

[4] Y. Qi and B. Weinberg, “Low-frequency energy deficit in electrolaryngeal speech,” *J. Speech Hear. Res.*, vol. 34, pp. 1250–1256, 1991.

[5] G. S. Meltzner, “Perceptual and acoustic impacts of aberrant properties of electrolaryngeal speech,” Ph.D. dissertation, Harvard-MIT Division of Health Sciences and Technology, Harvard Univ., Cambridge, MA, USA, 2003.

[6] R. Myrick and R. Yantorno, “Vocal tract modeling as related to the use of an artificial larynx,” in *Proc. IEEE 19th Annu. Northeast Bioeng. Conf.*, 1993, NJ, USA, pp. 212–214.

[7] K. Ma, P. Demirel, C. Espy-Wilson, and J. MacAuslan, “Improvement of electrolaryngeal speech by introducing normal excitation information,” in *Proc. Eurospeech*, Budapest, Hungary, 1999, pp. 323–326.

[8] K. M. Houston, R. E. Hillman, J. B. Kobler, and G. S. Meltzner, “Development of sound source components for a new electrolarynx speech prosthesis,” in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, 1999, pp. 2347–2350.

[9] K. Ooe, T. Fukuda, and F. Arai, “New type artificial larynx using PZT ceramics vibrator as sound source,” in *Proc. IEEE/ASME Int. Conf. Adv. Intell. Mechatron.*, 1999, pp. 114–119.

[10] G. Fant, J. Liljencrants, and Q. Lin, “A four-parameter model of glottal flow,” *Speech Trans. Lab. Q. Prog. Stat. Rep.*, vol. 4, pp. 1–13, 1985.

[11] J. O. Smith, “Physical modeling using digital waveguides,” *Comput. Music J.*, vol. 16, pp. 74–91, 1992.

[12] G. S. Meltzner, J. B. Kobler, and R. E. Hiillman, “Measuring the neck frequency response function of laryngectomy patients: implications for the design of electrolarynx devices,” *J. Acoust. Soc. Amer.*, vol. 114, pp. 1035–1047, 2003.

[13] B. H. Story, I. R. Titze, and E. A. Hoffman, “Vocal tract area functions from magnetic resonance imaging,” *J. Acoust. Soc. Amer.*, vol. 100, pp. 537–554, 1996.

[14] D. G. Childers, *Speech Processing and Synthesis Toolboxes*. New York, NY, USA: Wiley, 2000, app. 11.

[15] J. Clark and C. Yallop, *An Introduction to Phonetics and Phonology*. Oxford, U.K.: Blackwell, 1995, ch. 7.

[16] P. Ladefoged, *Vowels and Consonants: An Introduction to the Sounds of Language*. Oxford, U.K.: Blackwell, 2001, ch. 7.

Authors’ photographs and biographies not available at the time of publication.