Speech Communication 87 (2017) 31–39

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Acoustic influence of the neck tissue on Mandarin voiceless consonant production of electrolaryngeal speech

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

The Key Laboratory of Biomedical Information Engineering of the Ministry of Education, Department of Biomedical Engineering, School of Life Science and Technology, Xi'an Jiaotong University, No. 28, Xianning West Road, Xi' an 710049, PR China

ARTICLE INFO

Article history:
Received 9 November 2015
Revised 5 December 2016
Accepted 30 December 2016
Available online 4 January 2017

Keywords:
Electrolaryngeal speech
Neck frequency response function
Voiceless consonants

ABSTRACT

Lack of an appropriate voice source is the main reason for low intelligibility of the electrolaryngeal (EL) voiceless consonants. It is essential that the influence of neck tissue on EL voiceless consonant production is studied in order to design a suitable voice source. Firstly, the neck frequency response function was measured across a wide frequency range (100–20,000 Hz) to investigate the potential impact of the neck tissue on the EL voiceless consonants. The results show that the low-pass characteristic of the neck tissue distorts the spectral shape of the EL voiceless consonant, due to increased energy attenuation at higher frequencies (>2259 Hz). Then, a random and high-frequency energy-enhanced noise (HFEE-Noise) source was designed for EL voiceless consonant production. The results indicate that the HFEE-Noise source can compensate for the influence of the neck tissue, and shows promising potential for production of an EL voiceless consonant with a spectral shape that is closer to a natural voice. Furthermore, the results of perceptual experiments show that the HFEE-Noise source can effectively reduce the perceptual sonorization of the EL voiceless consonants, but shows limitations in increasing the perceptual discriminability of different voiceless consonants. Finally, some suggestions were proposed for the future design and implementation of an appropriate voice source to improve the EL voiceless consonant intelligibility.

1. Introduction

The electrolarynx (EL) is an electro-mechanical device designed for a patient who has lost his larynx or has a laryngeal disease and produces alaryngeal speech. The EL is placed against the neck to produce a mechanical sound which provides a substitute voice source for speech production. Although the EL has been widely used due to its ease of operation, it is still not satisfactory for daily communication due to the low intelligibility of EL speech (Hillman et al., 1998; Meltzner et al., 2005).

The failure of consonant (especially the voiceless consonants) production is the main deficiency of EL speech (Hillman et al., 1998; Weiss et al., 1979; Weiss and Basili, 1985). The principle role of EL is to provide an alternative voice source, which is very important to restore natural speech. Qi and Weinberg (1991) and Meltzner (2003) found that commercial EL voice sources contributed to an unnatural quality and low intelligibility of EL English speech. According to our previous work, the voice source of current commercial ELs is more suitable for production of voiced sounds than production of voiceless sounds in Mandarin speech (Wu et al., 2013), resulting frequently in perceptual sonorization and low intelligibility of the EL voiceless consonants (Xiao, 2012). Thus, it is imperative that an appropriate voice source for voiceless consonant production should be designed.

However, there have been no studies reporting improvements in the EL voiceless consonant source, especially in Mandarin speech. In our previous work, a supra-glottal voice source was proposed and proved to be effective in producing natural EL Mandarin vowels (Wu et al., 2013). However, the production process and acoustic properties of the voiceless consonant are different from those of the vowel. The voiceless consonant always has a higher frequency (>4000 Hz) and a wider band (2000–15,000 Hz) of energy concentration than the vowel (Cox, 2008). Therefore, the high-frequency acoustic features must be the key consideration in designing a voice source for EL voiceless consonant production.

The first step is to thoroughly investigate the potential impact of the neck tissue on the EL voice source. Several researchers have reported that the neck tissue acts as a low-pass filter (Meltzner et al., 2003; Wu et al., 2014), which will attenuate the high-frequency sound energy and distort the spectral shape of the EL voice source. In order to regain the high-frequency content, Meltzner et al. (2003) have approximated the neck frequency response function by a discrete linear filter and suggested compensation by inverse filtering of the EL driving signal. However, this research was

* Corresponding author.
E-mail addresses: liangwu@xjtu.edu.cn (L. Wu), mxwan@mail.xjtu.edu.cn (M. Wan).

http://dx.doi.org/10.1016/j.specom.2016.12.003
0167-6393/

focused only on the frequency range below 4000 Hz, and is therefore not able to provide useful guidance for improving the EL voice source for voiceless consonant production.

In this work, the acoustic influence of the neck tissue on the Mandarin voiceless consonant production of EL speech was studied to propose a preliminary design for an EL voiceless consonant voice source. Firstly, the neck frequency response function (NFRF) across a wide frequency range (100–20,000 Hz) was measured simply and accurately using a reflectionless tube, in order to evaluate the potential influence of neck tissue on EL voiceless consonant production. Then, a broadband random noise source with enhanced high-frequency energy (HFEE-Noise source) was designed for Mandarin voiceless consonant production to compensate for the influence of neck tissue. Finally, the spectral shapes and perception ratings of the voiceless consonants produced with the HFEE-Noise source were compared to those produced with a commercial EL source and from a normal voice, to evaluate the feasibility of the HFEE-Noise source and provide some suggestions for the future design and implementation of an EL voiceless consonant source.

2. Methods and experiments

2.1. Subjects

Five males and five females (average age of 25.7 ± 2.2 years) participated in the following experiments. All subjects were native Mandarin Chinese speakers, with no reported history of speech problems. These participants were familiar with EL speech production using a commercial EL.

2.2. NFRF measurement procedures

The NFRF was defined as the ratio of the spectrum of the estimated volume velocity that excites the vocal tract to the spectrum of the acceleration measured at the neck. The NFRF was measured using a reflectionless uniform tube, which in our previous work proved to be an effective method which was physically simple and resistant to noise (Wu et al., 2014).

Fig. 1(a) shows the NFRF measurement process. During the experiment, the subject was requested to sit in front of a reflectionless tube, hold a shaker (Brüel & Kjær, Model 4810, Skodsborgvej, Denmark) against his neck and connect his mouth to the tube through a detachable plastic mouthpiece. The shaker was a linear vibration transducer with an attached impedance head (KISTLER 8770A5, Winterthur, Switzerland) to measure the acceleration and a 3 cm diameter metal cap to provide a sufficient contact region with the neck. The reflectionless tube was made of steel, with a length of 190 cm and an inner diameter of 2.5 cm. The tube end was filled with a 90 cm polyurethane foam wedge to minimize acoustic reflections. An electret microphone (Knowles, Model WP-25993-D63, Itasca, IL) was fitted 30 cm from the mouth end to collect the pressure signal in the tube. To verify the reflectionless performance, the tube was directly connected to the B&K shaker which was driven by a broadband random noise. Then, the shaker vibration acceleration and the sound pressure in the tube were measured simultaneously and the frequency response of the tube was computed as the spectral ratio of the pressure signal to the vibration signal. Fig. 2 shows that the magnitude frequency response curve is flat enough (0.72 ± 0.65 dB) to neglect the influence of resonance in the tube.

Five Mandarin vowels (Pinyin *a, o, e, i, u*, as [ɑ], [ɔ], [ə], [i], [u] in IPA) were selected for NFRF measurement. In each trial, a broadband random noise was used to drive the shaker, and the shaker was placed on the right thyrohyoid membrane, superior to the thyroid rim, which is where the EL is commonly placed (Meltzner, 2003). The subjects were asked to produce each Mandarin vowel for 2–4 s, by just using the vocal tract structure without vocal fold vibration. Each vowel was repeated five times. During this, the pressure and the acceleration signals were synchronously recorded by a data acquisition system (BioPac MP150, Goleta, CA) and digitized at a sampling rate of 44,100 Hz with 16-bit quantization.

Similar to our previous work (Wu et al., 2014), the NFRF was computed here as follows: the coherence function between the acceleration and the pressure signal was firstly computed to confirm the assumption that the neck tissue is a linear time-invariant system. If more than 80% of the coherent functions of the data were above 0.8 between 100–20,000 Hz, the data was selected for NFRF analysis. Secondly, the NFRF was directly calculated as

$$N(f) = \frac{\Phi_{ap}(f)}{\Phi_{aa}(f)} = \frac{\mathcal{F}[\phi_{ap}(n)]}{\mathcal{F}[\phi_{aa}(n)]} \quad (1)$$

where $\Phi_{ap}(f)$ is the cross-spectral density of the acceleration and the pressure signal, computed as the Fourier transform of their cross-correlation $\phi_{ap}(n)$ and $\Phi_{aa}(f)$ is the power spectral density of the acceleration signal, calculated as the Fourier transform of the autocorrelation $\phi_{aa}(n)$. Finally, fifty NFRFs (10 subjects $\times$ 5 vowels) were measured, and each NFRF was averaged over the five trials. To minimize the effect of the vocal tract configuration on the NFRF estimation (Wu et al., 2014), the NFRF of each subject was calculated as the average NFRF of the five vowels.

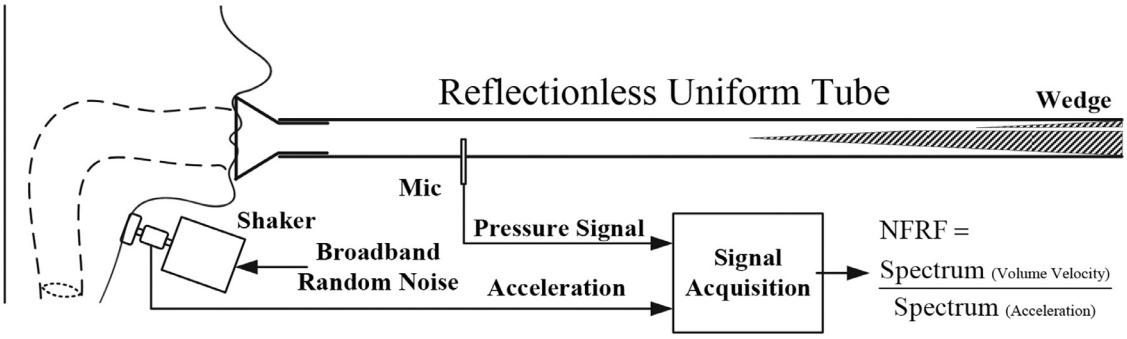
2.3. EL voiceless consonant production experiment

Fig. 1(b) shows the experimental procedures of EL voiceless consonant production. The subject held the same shaker in the same position to produce a Mandarin voiceless consonant. A dynamic microphone (Salar M9, Guangzhou, China) was mounted 5 cm in front of the mouth to collect the speech signal. The data acquisition system was used to synchronously record the speech signal from the microphone and the acceleration signal from the impedance.

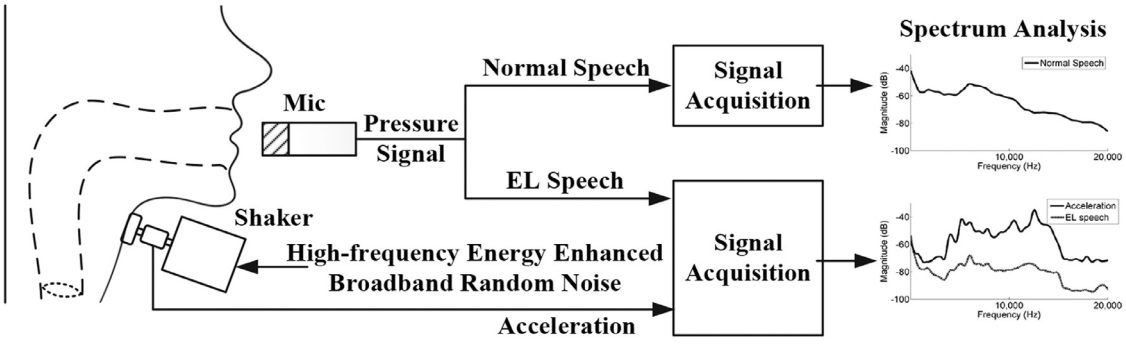
During the experiment, the shaker was driven by a designed waveform based on two aspects. Firstly, unlike vowels, a voiceless consonant is articulated by complete or partial closure of the vocal tract. Therefore, voiceless consonants are always characterized by a noise component other than periodic sound, and the acoustic energy is generally concentrated in a wide frequency range from 4000 Hz to 15,000 Hz (Cox, 2008; Jongman et al., 2000). This is particularly true for fricative or affricate consonants. Secondly, the NFRF (see Results section) shows that the neck tissue attenuates more high frequency energy than low frequency energy. Thus, the driving signal was designed as a broadband random noise (as shown in Fig. 3) to deliberately enhance the high-frequency components and suppress the low-frequency components without introducing any spectral features. The average magnitude in the 4000–15,000 Hz range is about 40 dB higher than that in other frequency ranges.

The ten Mandarin voiceless consonants selected in this experiment are listed in Table 1, which are commonly used voiceless initials in standard Chinese (Lee and Zee, 2003). In each trial, the subject was instructed to produce one voiceless consonant with the HFEE-Noise source and repeat it three times. To imitate EL speech production, the speaker was requested to hold his glottis closed during phonation. In addition, the voiceless consonants produced by a normal voice and the commercial EL source (Servox digital, Servona, Germany) were also collected separately by the data acquisition system with the microphone.

All the recordings were performed in a soundproof room. A one-second duration waveform was selected from the stable part of each recording to calculate the spectrum using a 2048-point



(a) Measurement of Neck Frequency Response Function



(b) Experiment of Electrolaryngeal Consonant Production

Fig. 1. Schematic diagrams of the experimental system and procedures. (a) Measurement of the neck frequency response function with a reflectionless uniform tube. (b) Experiment of EL consonant production with a high-frequency energy enhanced broadband random noise.

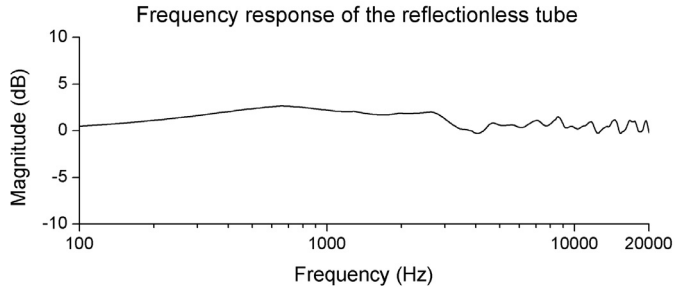


Fig. 2. Magnitude frequency response of the reflectionless tube. The maximum and minimum magnitudes are 2.64 dB and −0.29 dB, respectively.

Table 1
List of Mandarin voiceless consonants for EL production experiment.[a]

Affricate		Fricative
Unaspirated	Aspirated	
z [ts]	***c*** [tsʰ]	***s*** [s]
zh [tʂ]	***ch*** [tʂʰ]	***sh*** [ʂ]
j [tɕ]	***q*** [tɕʰ]	***x*** [ɕ]
		f [f]

[a] The bold letters represent the Pinyin, and the symbols in the brackets represent the corresponding transcription in the International Phonetic Alphabet (IPA).

discrete Fourier transform with a 50 ms moving Gaussian window. The voiceless consonants produced with the HFEE-Noise source and the commercial EL source were normalized for comparison purpose, in order to yield the same area under the spectrum curve

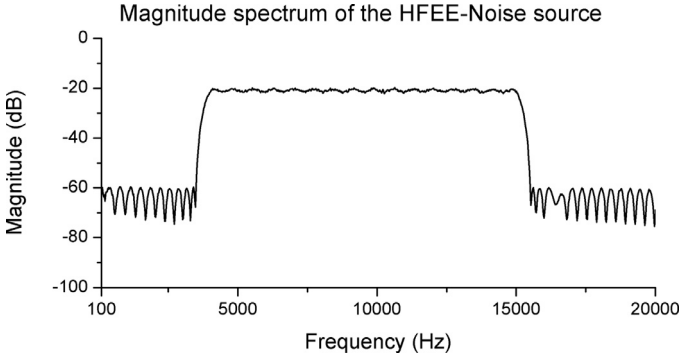


Fig. 3. Magnitude spectrum of the HFEE-Noise source. The average magnitude in the 4000–15,000 Hz range is −20.9 dB, and the average magnitude in other frequency ranges is −61.1 dB

as a normal voiceless consonant between 100 and 20,000 Hz. The following equation was used to normalize the spectrum:

$$P_{normlized}(f) = P_{EL}(f) + \frac{\int_{100}^{20000} P_{normal\_speech}(f)df - \int_{100}^{20000} P_{EL}(f)df}{19900} \quad (2)$$

where $P_{EL}(f)$ and $P_{normal\_speech}(f)$ are the spectra of the voiceless consonants produced with any EL voice source (the HFEE-Noise or the commercial EL source) and the normal voice, and $P_{normlized}(f)$ is the normalized spectrum. Finally, the neck influence on EL voiceless consonant production was evaluated based on a comparison of the normalized spectra.

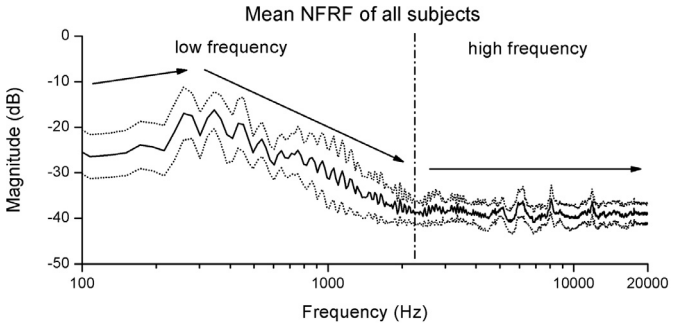


Fig. 4. Average NFRF (solid line) of all subjects with the standard deviations (dotted line). The lines with arrows illustrate the curve trend of the NFRF. The dash dotted line represents the stopband cutoff frequency at 2259 Hz.

2.4. Perceptual experiments

Ten listeners participated in the task, including five males and five females aged from 22 to 31 years (averaging 26.4 years). All listeners were Chinese native speakers and had no reports of any hearing or language disorders. The listeners were not experts on perceptual evaluation of speech, but they were familiar with EL speech.

An intelligibility test and a rating test were performed to obtain an overall perception evaluation of the EL voiceless consonants produced with the HFEE-Noise source. Two sets of listening samples were selected from the recordings of the production experiment, one male and one female. Each set contained 20 recordings (10 voiceless consonants $\times$ 2 sources).

For the intelligibility test, listeners were instructed to write down the phoneme heard using broad phonetic transcription, listening to the material as many times as needed. To avoid learning effect, the order of playback was randomly set. Then, voiceless consonant intelligibility was calculated using the average percentage of correct responses for all listeners. Additionally, the perceptual degree of sonorization was calculated using the average percentage of responses that was misidentified as voiced sound for all listeners.

For the rating test, the materials were regrouped to 10 sets according to different voiceless consonants. Each set contained four EL voiceless consonants (2 genders $\times$ 2 sources) and the corresponding normal speech. Listeners were asked to rate the perceptual similarity of the EL voiceless consonants to normal speech using a visual sort and rate (VSR) method (Granqvist, 2003; Chen et al., 2013). Finally, the average similarity rating of EL voiceless consonants for all listeners was calculated in the range from 0 to 1.

3. Results

3.1. Neck frequency response function curve

3.1.1. General description

The mean NFRF and standard deviations for all subjects between 100–20,000 Hz is shown in Fig. 4. Firstly, the NFRF resembles a low-pass filter in its overall shape, and the cutoff frequency of the stop band is approximately 2259 Hz. Therefore, the NFRF curve was divided into a low frequency section (<2259 Hz) and a high frequency section. In the low frequency section, the NFRF curve has a maximum peak at 223.9 ± 44.5 Hz, and then rolls off at a slope of -7.8 ± 1.6 dB/octave. In the high frequency section, the NFRF flattens out as it reaches 20,000 Hz, with only several minor peaks. Additionally, the standard deviation (2.6 ± 0.8 dB) reflects individual differences of the NFRF. The standard deviations of low frequency and high frequency sections were 4.6 ± 1.2 dB and 2.4 ± 0.3 dB, respectively.

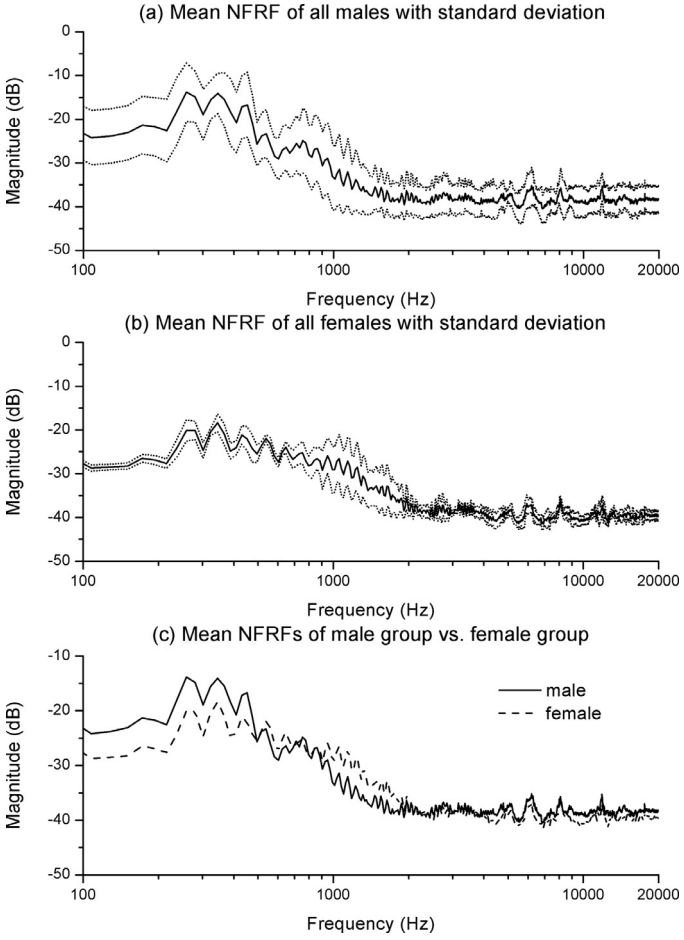


Fig. 5. Average NFRFs (solid line) and the standard deviations (dotted line) for (a) the male group and (b) the female group. (c) Comparison of the average NFRFs between the male group (solid line) and the female group (dash line).

3.1.2. Male group vs. female group

Fig. 5 plots the average NFRFs of different gender groups. The results show an obvious difference in the NFRF shape for the male and female groups in the low frequency section. Firstly, the average peak frequency of the NFRFs of the male subjects was 206.7 ± 47.2 Hz, which was a little lower than 241.2 ± 38.5 Hz of the female subjects. Secondly, the average descending slopes of the NFRFs were -8.9 ± 1.2 dB/octave for the male subjects and -6.6 ± 0.7 dB/octave for the female subjects, respectively. Thirdly, the average stopband cutoff frequency of the NFRFs of the male subjects was 1809 Hz, which was lower than that of the female subjects (2453 Hz). In the high frequency section, although the NFRF curves of both groups were flat, the standard deviation of the male subjects (3.2 ± 0.5 dB) was a little larger than that of the female subjects (1.3 ± 0.4 dB, T-test, $p < 0.001$).

3.1.3. Low frequency vs. high frequency

The average magnitudes of low frequency and high frequency sections were calculated as illustrated in Fig. 6, in order to compare the sound energy attenuation by the neck tissue. The average stopband cutoff frequency of all subjects (2259 Hz) was selected to distinguish between the two frequency sections. For all subjects, the average magnitudes of low frequency and high frequency sections were -30.01 ± 3.45 dB and -39.86 ± 2.23 dB, respectively, with a maximum magnitude difference of 14.34 dB For different

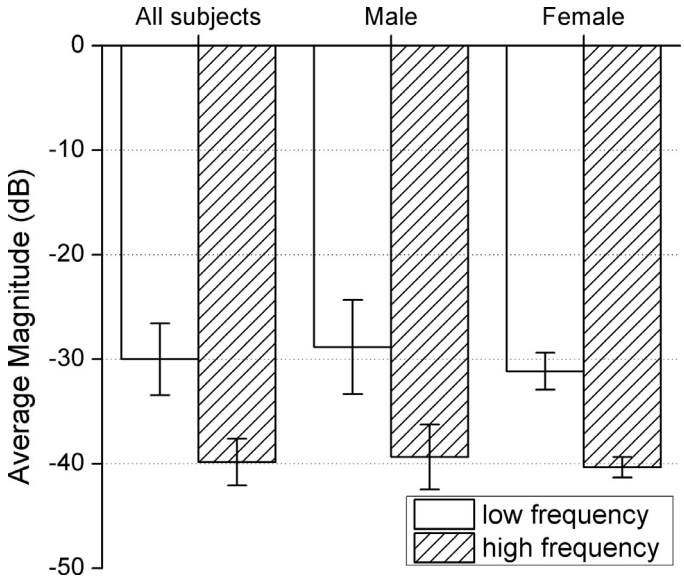


Fig. 6. Average magnitudes of different frequency ranges. The low frequency refers to the range from 100 Hz to 2259 Hz, and the high frequency refers to the range from 2259 Hz to 20,000 Hz.

gender groups, the average magnitudes were not significantly different (*T*-test, $p = 0.68 > 0.05$) in the high frequency section. However, in the low frequency section, the average magnitude of male subjects (-28.84 ± 4.51 dB) was higher than that of female subjects (-31.17 ± 1.77 dB, $p = 0.009 < 0.05$). In addition, the standard deviations for male subjects were larger than for female subjects in both the low and high frequency sections.

3.2. Spectrum of EL voiceless consonants

3.2.1. General description

The average spectrum of the EL voiceless consonants produced with the HFEE-Noise source for all subjects is shown in Fig. 7(a), which compares the spectrum with the average spectrum of the normal speech. The spectrum of the HFEE-Noise source is also plotted to represent the actual transmission into the neck tissue. The results show that the magnitude of the EL voiceless consonant spectrum was 27.55 dB lower on average than that of the HFEE-Noise source spectrum, indicating energy attenuation across the whole frequency range. The EL voiceless consonant spectrum also shows frequency-related differences with the normal speech spectrum. As illustrated in Fig. 7(b), within the frequency ranges of 100–4000 Hz and 4001–10,000 Hz, the average magnitudes of the EL voiceless consonant spectrum were lower than those of the normal speech spectrum by an average of 7.69 dB and 4.65 dB (*T*-test, $p < 0.001$), respectively. In the 10,001−15,000 Hz and 15,001–20,000 Hz ranges, the average magnitudes of the EL voiceless consonant spectrum were 7.08 dB (*T*-test, $p < 0.001$) and 1.62 dB (*T*-test, $p = 0.062 > 0.05$) higher than those of the normal speech spectrum.

3.2.2. Difference between subject groups

Fig. 8 plots the average spectra of the voiceless consonants for the male and female groups produced with (a) the HFEE-Noise source (b) and normal voice. The results indicate an obvious difference between the EL voiceless consonant spectra of males vs. females. The spectrum magnitude of speech produced by male subjects was lower than that for female subjects by 7.83 ± 3.20 dB (*T*-test, $p < 0.001$) across the whole frequency range, indicating more energy attenuation for male subjects. The same difference was also

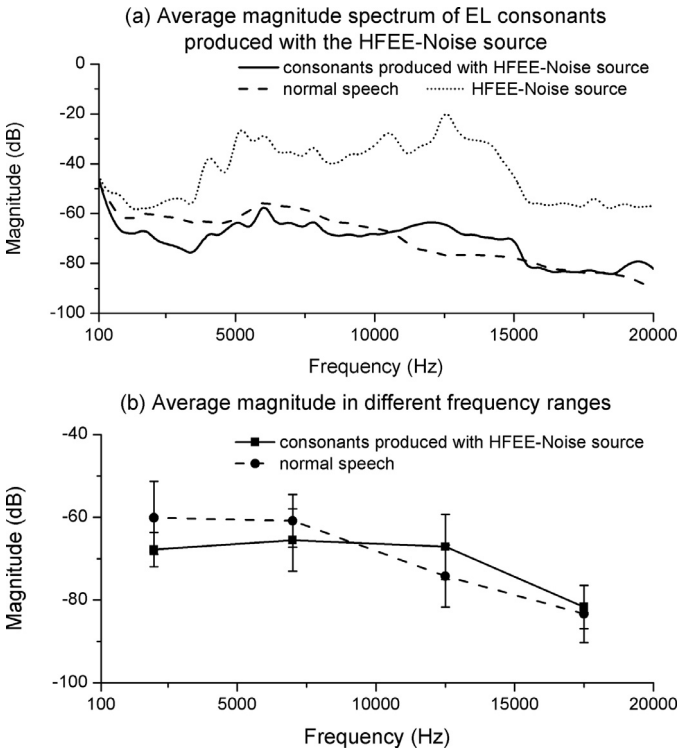


Fig. 7. (a) Mean value of the normalized spectra of the consonants produced with the HFEE-Noise source (solid line) for all subjects. Dash line refers to the average spectrum of the normal speech, and dotted line represents the spectrum of the actual vibration signal transmitting into the neck. (b) Average magnitudes in different frequency ranges. Four frequency ranges are 100–4000 Hz, 4001–10,000 Hz, 10,001–15,000 Hz, and 15,001–20,000 Hz, respectively. The x-coordinate of each point is the middle value of the corresponding frequency range. The error bar represents the individual difference.

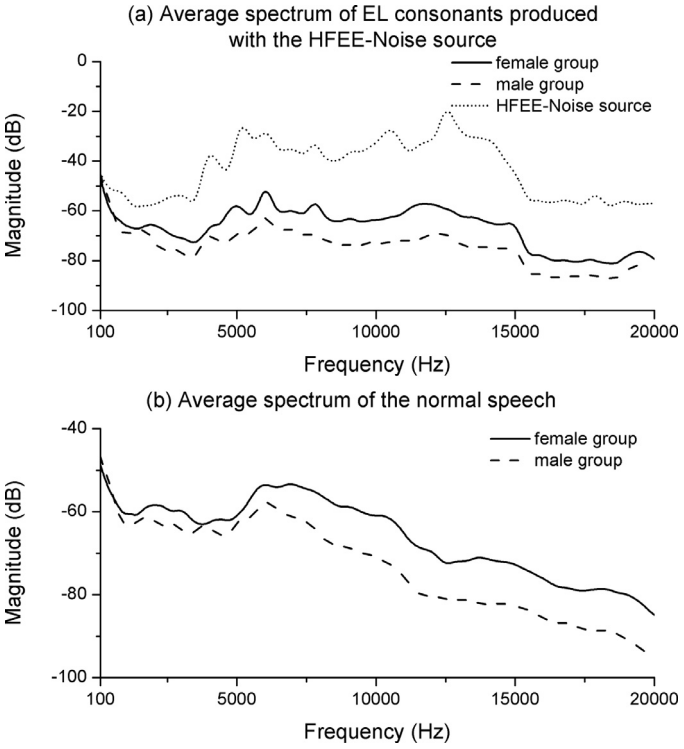


Fig. 8. Spectra of the consonants for different gender groups produced with (a) the HFEE-Noise source and (b) the normal voice. Solid line represents the female group, and dash line represents the male group. Dotted line in (a) is the spectrum of the HFEE-Noise vibration source.

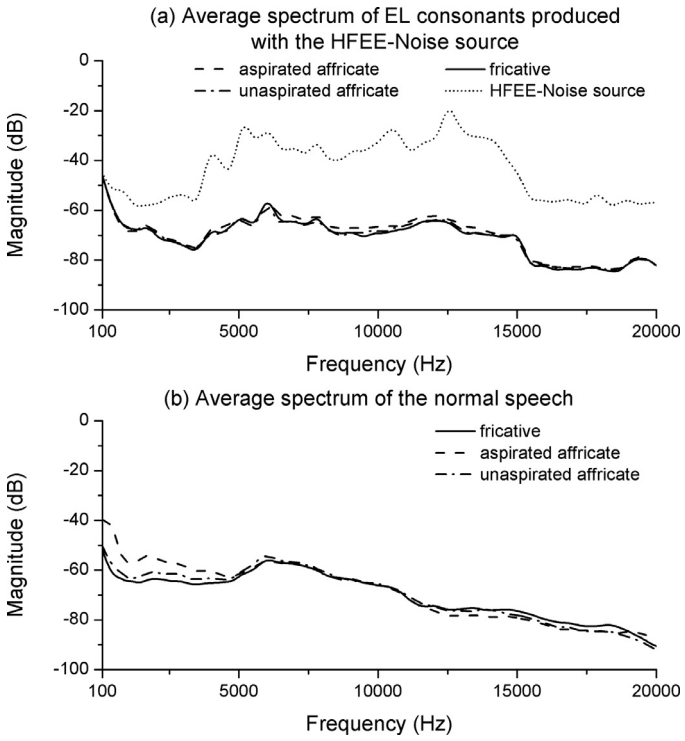


Fig. 9. Spectra of the different consonants produced with (a) the HFEE-Noise source and (b) the normal voice. Solid line refers to the fricative, and dash line refers to the aspirated affricate, and dash-dot line represents the unaspirated affricate. Dotted line in (a) is the spectrum of the HFEE-Noise vibration source.

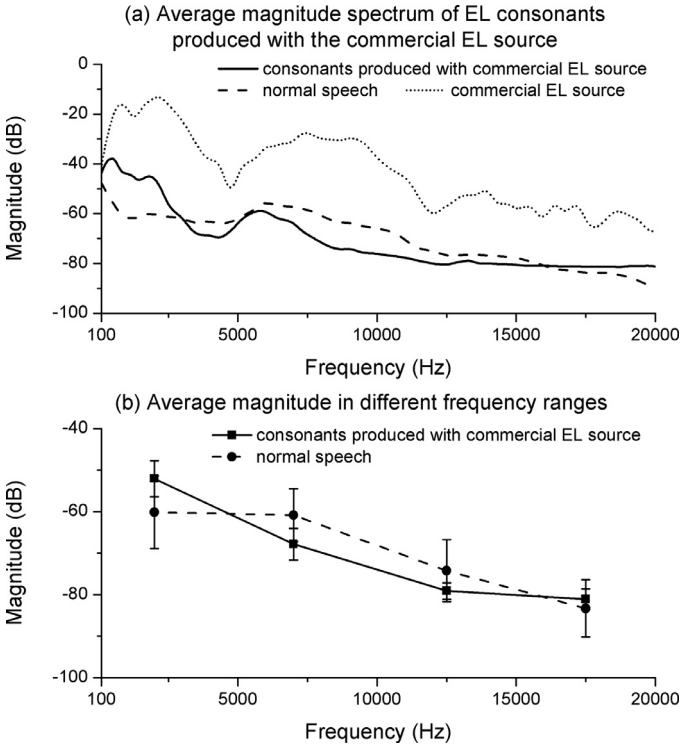


Fig. 10. (a) Mean value of the normalized spectra of the consonants produced with the commercial EL source (solid line) for all subjects. Dash line refers to the average spectrum of the normal speech, and dotted line represents the spectrum of the commercial EL source. (b) Average magnitudes in different frequency ranges.

found in the spectra of normal speech. The average magnitude of speech produced by male subjects was 7.48 ± 3.51 dB lower than that of female subjects (T-test, $p < 0.001$).

3.2.3. Difference across voiceless consonant types

Fig. 9 plots the average spectra of the voiceless consonants for different voiceless consonant types produced with (a) the HFEE-Noise source and (b) normal voice. The results showed a slight difference in the EL voiceless consonant spectra of different voiceless consonant types. The magnitude differences of the EL voiceless consonant spectra between any two voiceless consonant types were lower than 3.14 dB However, the spectra for normal speech were different than the voiceless consonant types at frequencies below 5000 Hz. The spectrum magnitude of the aspirated affricate was higher than those of the unaspirated affricate and the fricative by an average of 5.30 dB and 7.46 dB (T-test, $p < 0.001$). Above 5000 Hz, there was no obvious variance in spectra between any two voiceless consonant types with a maximum magnitude difference of 3.91 dB.

3.2.4. Commercial EL voiceless consonant spectrum

The average spectrum of the EL voiceless consonants produced with the commercial EL source was measured and compared with the normal speech as shown in Fig. 10(a). First, the energy attenuation in the whole frequency range was similar to the case of EL voiceless consonant production with the HFEE-Noise source. The average spectrum of the commercial EL voiceless consonants was lower than that of the commercial EL source by the average magnitude of 27.95 dB. Secondly, the spectrum of the commercial EL voiceless consonants was also different from that of the normal speech. The average magnitude of the commercial EL voiceless consonants was 8.09 dB higher than that of normal speech below 4000 Hz, but 7.00 dB and 4.89 dB lower in the 4001–10,000 Hz and 10,001–15,000 Hz ranges (T-test, $p < 0.001$), respectively.

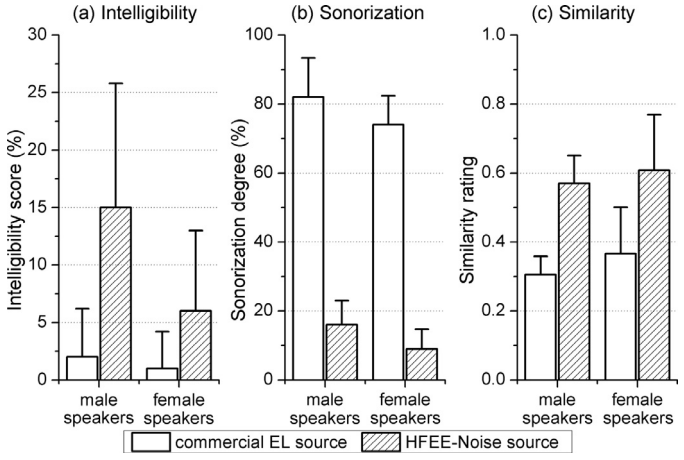


Fig. 11. Perceptual evaluation of the EL voiceless consonants produced with the HFEE-Noise source and the commercial EL source. (a) Intelligibility scores. (b) Perceptual sonorization degrees. (c) Similarity ratings.

3.3. Perceptual evaluation of EL voiceless consonants

Fig. 11 shows the perceptual results of the EL voiceless consonants produced with the HFEE-Noise source and the commercial EL source. The average voiceless consonant intelligibilities of the HFEE-Noise source were 15.0% for the male speakers and 6.0% for the female speakers, which were a little higher than those of the commercial EL source (2.0% for the male speakers, T-test, $p = 0.002 < 0.05$, and 1.0% for the female speakers, $p = 0.046 < 0.05$). For the degree of perceptual sonorization, a higher value indicates that a voiceless consonant is more easily misidentified as a voiced one. The average perceptual sonorization degree of the HFEE-Noise source were $16.0 \pm 7.0\%$ for the male

speakers and 9.0 ± 5.7% for the female speakers, which were significantly lower than those of the commercial EL source (82.0 ± 11.4% for the male speakers and 74.0 ± 8.4% for the female speakers, *T*-test, $p < 0.001$). Finally, for the similarity rating, a higher value indicates that the sample is perceived as being closer to normal speech. The average similarity ratings of the HFEE-Noise source were 0.570 ± 0.081 for the male speakers and 0.608 ± 0.161 for the female speakers, which were significantly higher than those of the commercial EL source (0.305 ± 0.054 for the male speakers, *T*-test, $p < 0.001$, and 0.366 ± 0.135 for the female speakers, $p = 0.002 < 0.05$).

4. Discussion

4.1. Influence of the neck tissue on EL voiceless consonant production

In this work, the NFRF was easily obtained using a reflectionless tube, which was proved to be an effective way to measure the sound transmission characteristics of the neck tissue at frequencies below 5000 Hz (Wu et al., 2014). In Fig. 4, the average NFRF curve of ten subjects shows a very similar shape in the 100–5000 Hz range as obtained by previous results (Meltzner et al., 2003; Wu et al., 2014; Norton and Bernstein, 1993). This outcome demonstrated that our procedure was feasible and the results were reliable. Furthermore, the flat frequency response of the reflectionless tube in Fig. 2 indicates strong performance in the high frequency ranges (>5000 Hz). Therefore, the NFRFs measured here can be used to reflect the acoustic influence of the neck tissue across a wide frequency range from 100 to 20,000 Hz.

In Fig. 4, it was again confirmed that the low-pass characteristic was key factor influencing the role of the neck tissue in sound transmission. In the low frequency section (<2259 Hz), the pass band was lower than ~223.9 Hz with an average magnitude of −21.93 dB, and the transitional band was not steep but wide (~2000 Hz) with a slope of −7.8 dB/octave. In contrast, the stop band was stable at −39.86 dB from 2259 to 20,000 Hz. These results indicate that the neck tissue may influence the EL voiceless consonant production in two ways. Firstly, the acoustic energy of the voiceless consonant is generally distributed in the 4000–15,000 Hz range (Cox, 2008; Jongman et al., 2000). The high energy attenuation at higher frequencies will distort the spectral shape of the EL voiceless consonant, thus causing confusion in perception of different voiceless consonants. Secondly, the energy attenuation at low frequencies is 9.85 dB lower than at high frequencies. Therefore, the acoustic energy at low frequencies will be enhanced by some degree, thus causing a perceptual sonorization of the EL voiceless consonant, although the flat curve of the NFRF at high frequencies still means that the neck tissue will not introduce unwanted spectral peaks into the spectrum of the EL voiceless consonant.

For different subjects, Fig. 5 shows a remarkable difference in the NFRF due to gender in the low frequency section but not in the high frequency section. Furthermore, the magnitude difference of male subjects between the low and high frequency ranges (10.53 dB) was larger than that of female subjects (9.17 dB). These results are likely due to the physiological structure difference of the larynx between the male and female groups (Wu et al., 2014). This outcome indicates that the EL voiceless consonants produced by male subjects may be more easily perceived as voiced consonants than if produced by a female subject.

In conclusion, the neck tissue potentially impacts the spectral shape and auditory perception of the EL voiceless consonants due to large energy attenuation at high frequencies. Therefore, high-frequency energy compensation is necessary when designing a voice source to improve the perceptual intelligibility of the EL voiceless consonants.

4.2. Feasibility of the HFEE-Noise source in EL voiceless consonant production

Based on this analysis of the influence of the neck tissue, we designed a broadband random noise source with enhanced high-frequency energy (HFEE-Noise source), to study the feasibility of high-frequency energy compensation to improve the EL voiceless consonant production. The HFEE-Noise source was not designed for any specific consonant. A magnitude difference of 40 dB in the driving signal was designed to be equal to the high-frequency energy attenuation of the neck tissue, aiming to compensate for the influence of the neck tissue and enhance the acoustic energy of the EL voiceless consonants. However, the actual vibration source for the EL voiceless consonant production was distorted due to nonlinear vibrations of the shaker. The actual high-frequency energy compensation was an average of 23 dB. Even so, the results are still meaningful and reflect the effect of the high-frequency energy compensation on EL voiceless consonant production.

Fig. 7 shows the average spectrum for the EL voiceless consonants produced with the HFEE-Noise source. Although the spectrum of the EL voiceless consonants was different from that of normal speech, two spectral features still show the effectiveness of the high-frequency energy compensation. Firstly, the spectrum magnitude of the EL voiceless consonants at high frequencies (>4000 Hz) was no lower than that of normal speech, and was higher than that of the EL voiceless consonants at low frequencies (<4000 Hz). This result indicates that the HFEE-Noise source can preserve or enhance the acoustic energy of the EL voiceless consonants at high frequencies. Secondly, the magnitude of the EL voiceless consonants at low frequencies was lower than that of normal speech. These results indicate that the HFEE-Noise source was able to relatively reduce the low-frequency energy of the EL voiceless consonants. Therefore, it is feasible to compensate for the influence of the neck tissue by enhancing the high-frequency energy in the EL source.

For different gender groups, the spectra of the EL voiceless consonants produced with the HFEE-Noise source were significantly different. The spectrum magnitude of male subjects was lower than that of female subjects across the entire frequency range. Since the NFRFs of the male and female groups were not different at high frequencies, these differences of EL voiceless consonant spectra may be attributed to differences in the vocal tract resonance. This hypothesis can be confirmed by examining the same spectral differences between the male and female groups in normal speech, as shown in Fig. 8(b). Thus, the spectral shape differences of different gender groups will not be influenced by the neck tissue and will be reserved for the EL voiceless consonants produced with the HFEE-Noise source.

For different voiceless consonant types, there was no difference in the spectra of the EL voiceless consonants produced with the HFEE-Noise source. This result suggested that the contribution of the vocal tract resonance to EL voiceless consonant production was not large enough to distinguish different voiceless consonant types. Particularly at low frequencies, the spectrum magnitude of normal speech was different for different voiceless consonant types. Thus, the spectral shapes of different voiceless consonant types cannot be produced through the HFEE-Noise source or vocal tract resonance.

The spectrum of the EL voiceless consonants produced with the HFEE-Noise source was not the same as that of normal speech, but was still closer to a natural voice than that produced with the commercial EL source. Firstly, the average magnitude of the EL voiceless consonants produced with the HFEE-Noise source at high frequencies (>4000 Hz) was 4.42 dB higher than that produced with the commercial EL source, and closer to that of normal speech. Secondly, the average magnitude of the EL voiceless con-

sonants produced with the HFEE-Noise source at low frequencies (<4000 Hz) was 13.44 dB lower than that produced with the commercial EL source, and closer to that of normal speech. Thirdly, the low-frequency magnitude of the EL voiceless consonants produced with the commercial EL source was higher than the high-frequency magnitude, while it was the opposite in the case of the HFEE-Noise source. All these differences indicate that the EL voiceless consonants produced with the HFEE-Noise source had more noticeable high-frequency spectral characteristics than that produced with the commercial EL source, which will be helpful to reduce the perceptual sonorization and increase the intelligibility of the EL voiceless consonants.

The perceptual experiments in this study evaluated the contribution of the HFEE-Noise source on the perception of the EL voiceless consonants. Firstly, the intelligibility and similarity of the EL voiceless consonants produced with the HFEE-Noise source were higher than those with the commercial EL source. There results indicate that the HFEE-Noise source can improve the perceptual quality of the EL voiceless consonants. It is thought that this improvement is mainly related to the low perceptual sonorization of the EL voiceless consonants as shown in Fig. 11, which demonstrates that the HFEE-Noise source can effectively reduce the misperception of voiceless consonants as voiced ones due to the increased high-frequency energy and relatively decreased low-frequency energy in the spectrum. Secondly, although the perceptual sonorization of the EL voiceless consonants was highly improved, the intelligibility and similarity of the EL voiceless consonants produced with the HFEE-Noise source only show limited improvements which were still far lower than normal speech. This outcome can be mainly explained by the high perceptual confusion of different EL voiceless consonants, which were 82.14% for the male speakers and 93.41% for the female speakers in the case of the HFEE-Noise source and 88.89% for the male speakers and 96.15% for the female speakers in the case of the commercial EL source. These results showed that the HFEE-Noise source cannot effectively improve the perception discrimination of different EL voiceless consonants, which is consistent with the conclusions drawn from spectrum comparisons of different voiceless consonant types. Furthermore, the perceptual evaluation in this work was based on single phonemes. For a running speech, the EL consonant intelligibility may be higher due to the context information (such as stimulus length, coarticulation, etc. Diehl et al., 2004), but this effect needs further evaluation in daily EL communication.

4.3. Suggestions on voice source design for EL voiceless consonant production

In this work, a HFEE-Noise source was designed to compensate for the influence of the neck tissue on EL voiceless consonant production. However, beyond that, the contribution of the HFEE-Noise source to the spectral shape and the auditory perception of the EL voiceless consonants can provide some further suggestions for developing an appropriate voice source for EL voiceless consonant production.

(1) A noise component may be a better choice than a periodic source, since periodic sources such as the current commercial EL source have concentrated energy at low frequencies, which will mask the high-frequency characteristics of the EL voiceless consonants. Turbulence is generally considered to be the common source for most voiceless consonants during normal speech production (Narayanan and Alwan, 2000).
(2) A high-frequency energy enhancement is necessary, not only to compensate for the energy attenuation of the neck tissue, but also to increase the perceptual discriminability of the voiceless consonants from the voiced consonants.
(3) The EL consonant voice source should be designed specially for different voiceless consonant phonemes, because the vocal tract resonance is not strong enough to produce a satisfactory acoustic and perceptual distinction of different voiceless consonant phonemes.

However, this work was done by healthy subjects but not laryngectomized patients. The physiological structure differences, especially a shorter vocal tract length resulted from the laryngectomy, will distort the frequency features in the EL consonants as in the EL vowels (Wu et al., 2013). So the effect of the shorten vocal tract should be considered in future design of a consonant voice source.

4.4. Implementation and application

On account of the big difference between the vowel production and consonant production, the proposed voiceless consonant source is not designed and may be not suitable for the EL vowel production. The reduced low-frequency energy may produce weaken formant features and affect perceptual distinguishability of different EL vowels. Thus, the proposed voiceless consonant source should be used together with the vowel source in real application by switch control according to different phonemes. The EMG technology is seemed to be a possible way for EL source control because the EMG has been investigated in the syllable and phoneme recognition and classification (Bu et al., 2005; Yau et al., 2008; Lopez-Larraz et al., 2010) and has been used to model coarticulation in the continuous speech recognition (Schultz and Wand, 2010). Furthermore, the fact that neck trap muscle EMG precedes voice by 70 or 120 ms (Atkinson, 1978) is an advantage to achieve a real-time control. However, realizing an accurate control of the EL sources is not easy work and needs to be further studied and evaluated in practical use.

5. Conclusion

The average NFRF of ten subjects indicates that the neck tissue has a potential impact on EL voiceless consonant production. The high-frequency energy was attenuated much more than the low-frequency energy, which will distort the spectral shapes of the EL voiceless consonant and increase the perceptual sonorization. However, it was feasible to compensate for the neck tissue influence using a high-frequency energy enhancement in the EL voice source, producing an EL voiceless consonant with a closer spectral shape and auditory perception to a natural voice. In particular, the misperception of an EL voiceless consonant as a voiced consonant was greatly improved by compensating for the high-frequency energy in the voice source, although the intelligibility of the EL voiceless consonant cannot be increased significantly by only a uniform energy compensation for all voiceless consonant production. Therefore, a random noise with consonant-specific high-frequency energy compensation may be an appropriate voice source for EL voiceless consonant production. Our future work will synthesize different voice sources to improve the intelligibility of different voiceless consonants.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 11404256, 11274250, and 61271087) and Project funded by China Postdoctoral Science Foundation (No. 2014M552457), and by the Fundamental Research Funds for the Central Universities (No. xjj2014057).

References

Atkinson, J, 1978. Correlation analysis of the physiological factors controlling fundamental voice frequency. J. Acoust. Soc. Am. 63, 211–222.

Bu, N., Tsuji, T., Arita, J., Ohga, M., 2005. Phoneme classification for speech synthesiser using differential EMG signals between muscles. In: Engineering in Medicine and Biology Society, 27th Annual International Conference of the IEEE, pp. 5962–5966.

Chen, G., Garellek, M., Kreiman, J., Gerratt, B.R., Alwan, A., 2013. A perceptually and physiologically motivated voice source model. Interspeech 2013, 2001–2005.

Cox, F., 2008. Speech Acoustics: Consonant Acoustics. Centre for Language Sciences, Department of Linguistics, Macquarie University. http://clas.mq.edu.au/speech/acoustics/consonants/index.html (accessed on 11.9.2016).

Diehl, R.L., Andrew, J.L., Lori, L.H., 2004. Speech perception. Annu. Rev. Psychol. 55, 149–179.

Granqvist, S., 2003. The visual sort and rate method for perceptual evaluation in listening tests. Logoped. Phonatr. Vocol 28, 109–116.

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. Research Speech-Language Pathologists. Department of Veterans Affairs Laryngeal Cancer Study Group. Ann. Otol. Rhinol. Laryngol. Suppl. 172, 1–27.

Jongman, A., Wayland, R., Wong, S., 2000. Acoustic characteristics of English fricatives. J. Acoust. Soc. Am. 108, 1252–1263.

Lee, W., Zee, E., 2003. Standard Chinese (Beijing). J. Int. Phon. Assoc. 33, 109–112.

Lopez-Larraz, E., Mozos, O.M., Antelis, J.M., Minguez, J., 2010. Syllable-based speech recognition using EMG. In: Engineering in Medicine and Biology Society, 2010 Annual International Conference of the IEEE, pp. 4699–4702.

Meltzner, G.S., 2003. Perceptual and Acoustic Impacts of Aberrant Properties of Electrolaryngeal Speech. Harvard-MIT Division of Health Sciences and Technology, Harvard University, Cambridge, MA, USA.

Meltzner, G.S., Kobler, J.B., Hillman, R.E., 2003. Measuring the neck frequency response function of laryngectomy patients: implications for the design of electrolarynx devices. J. Acoust. Soc. Am. 114, 1035–1047.

Meltzner, G.S., Hillman, R.E., Heaton, J.T., Houston, K.M., Kobler, J.B., Qi, Y., 2005. Electrolaryngeal speech: the state of the art and future directions for development. In: Contemporary Consideration in the Treatment and Rehabilitation of Head and Neck Cancer: Voice, Speech, and Swallowing, pp. 571–590.

Narayanan, S., Alwan, A., 2000. Noise source models for fricative consonants. IEEE Trans. Speech Audio Process. 8, 328–344.

Norton, R.L., Bernstein, R.S., 1993. Improved Laboratory Prototype Electrolarynx (LAPEL): using inverse filtering of the frequency response function of the human throat. Ann. Biomed. Eng 21, 163–174.

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

Schultz, T., Wand, M., 2010. Modeling coarticulation in EMG-based continuous speech recognition. Speech Commun. 341–353.

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.

Weiss, M.S., Basili, A.G., 1985. Electrolaryngeal speech produced by laryngectomized subjects: perceptual characteristics. J. Speech Hear. Res. 28, 294–300.

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

Wu, L., Xiao, K., Dong, J., Wang, S., Wan, M., 2014. Measurement of the sound transmission characteristics of normal neck tissue using a reflectionless uniform tube. J. Acoust. Soc. Am. 136, 350–356.

Xiao, K., 2012. The Defects in the Consonant Production of Electrolaryngeal Speech Thesis. Xi'an Jiaotong University, Shaanxi, China.

Yau, W.C., Arjunan, S.P., Kumar, D.K., 2008. Classification of voiceless speech using facial muscle activity and vision based techniques. In: TENCON 2008, IEEE Region 10 Conference, pp. 1–6.