Speech Communication 123 (2020) 26–34

Contents lists available at ScienceDirect

Speech Communication

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

Aoustical and perceptual characteristics of mandarin consonants produced with an electrolarynx

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

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

ARTICLE INFO

Keywords:
Mandarin speech
EL consonants
Acoustic characteristics
Perceptual intelligibility

ABSTRACT

The electrolarynx (EL) is an electromechanical device that enables patients to produce voice following the surgical removal of their larynx. The purpose of this study is to understand the acoustic and perceptual characteristics of Mandarin consonants produced by EL speakers. First, the acoustic characteristics (including speech intensity, consonant duration, spectral peak, and *F2* onset) of Mandarin EL consonants are investigated by comparing the EL and normal consonants. Then, a perceptual evaluation of EL consonants is conducted to identify the relationship between acoustical characteristics and perceptual intelligibility. The results suggest three consonant confusion types are mainly responsible for the poor intelligibility of Mandarin EL consonants: (1) the "unaspirated-for-aspirated" confusion caused by the significantly shortened voice onset time of aspirated consonants; (2) the "voiced-for-voiceless" confusion caused by the continuous pulsing of the EL device and low consonant intensity; and (3) the "perceptual omission" caused by low intensity of consonants and / or consonant omission. The results obtained are promising and potential for further improvements in Mandarin EL speech intelligibility.

1. Introduction

Each year, more than 25,000 patients are diagnosed with laryngeal cancer in China (Chen et al., 2016). Laryngectomy remains a widely-used treatment method for patients diagnosed with advanced-stage laryngeal carcinoma. However, laryngectomy results in the inability to produce laryngeal voice due to removal of the larynx. The electrolarynx (EL) is an electromechanical device that enables patients to produce a voice even though their larynx has been surgically removed. Previous studies have indicated that more than half of laryngectomees use an EL up to two years post-laryngectomy due to a number of advantages, including its ease of learning and operation, and continuous output (Hillman et al., 1998). However, patients using ELs for communication are still limited in their communication due to the unnatural speech produced, thereby leading to low intelligibility (Kaye et al., 2017; Liu and Ng, 2007; Sluis et al., 2018; Verkerke and Thomson, 2014).

A number of speech characteristics have been previously reported to affect EL speech quality and intelligibility. Significantly low-frequency spectral energy deficit contributes to unnatural EL speech quality and low intelligibility (Qi and Weinberg, 1991; Meltzner, 2005). Higher fundamental frequencies have been shown to result in poorer intelligibility for English-speaking laryngectomees (Nagle et al., 2012). Furthermore, a lack of frequency variation has been shown to result in poorer intelligibility for tone-based languages, such as Thai and Cantonese (Gandour et al., 1988; Liu et al., 2006; Ng et al., 1998). Noise leakage from the EL also contributes to some of the unnatural characteristics of EL speech, again leading to poor intelligibility (Meltzner and Hillman, 2005; Meltzner et al., 2001). Shortened consonants are also an important factor that leads to low intelligibility of EL consonants (Hewlett et al., 1997).

Based on these shortcomings, it has been shown that EL speech quality can be enhanced by compensating for low-frequency energy deficits (Pandey and Basha, 2010), controlling EL fundamental frequency (Saikachi et al., 2009; Wan et al., 2012; Wang et al., 2018), reducing noise leakage (Basha and Pandey, 2012; Niu et al., 2003; Xiao et al., 2018), modifying the EL voice source (Espy-Wilson et al., 1998; Ng et al., 2014) and compensating for the abnormal vocal tract characteristics (Wu et al., 2013). While previous attempts have improved the acceptability of EL speech, only limited improvement of the intelligibility of EL speech was achieved. This is because the previous efforts mainly focused on the improvement of vowels in EL speech and largely ignored consonants. Wu et al. (2013) showed that current EL voice sources produce more intelligible vowels than consonants. However, consonants are indispensable for speech meaning discrimination (Owren and Cardillo, 2006). Consonant confusion in EL consonants can therefore cause a significant reduction in EL speech intelligibility

* Corresponding author.
E-mail address: liangwu@xjtu.edu.cn (L. Wu).

https://doi.org/10.1016/j.specom.2020.06.004
Received 2 April 2018; Received in revised form 21 January 2020; Accepted 28 June 2020
Available online 30 June 2020

(Weiss et al., 1979). Therefore, the purpose of the present study is to understand the acoustic and perceptual characteristics of Mandarin consonants produced by EL speakers.

1.1. Acoustic and perceptual characteristics of el consonants

Several acoustic characteristics have been demonstrated to be closely related with consonant perceptions. Consonants produced with larger airflows are usually characterized by larger intensity and longer consonant duration (Faulkner and Rosen, 1999; Shigeki et al., 2006), while the spectral peak (defined as the highest-amplitude peak of the FFT spectrum) is proportional to the airflow and inversely proportional to the diameter of constriction (Narayanan and Alwan, 2000). With the place of articulation moving to the glottis, the F2 onset (defined as the starting frequency of the second formant at the consonant-vowel boundary of voiceless consonants) gradually increases (Jongman et al., 2000; Maniwa et al., 2009; Sussman et al., 1991). Therefore, in this study, intensity and consonant duration were selected for acoustic analysis of consonants with different manners of articulation, while the spectral peak and *F2* onset were selected for acoustic analysis of consonants with different places of articulation.

So far, there have been several studies revealing that EL consonants have shorter duration than normal consonants, including shortened voice onset time (VOT, the time interval between the onset of the noise burst and the onset of the following vowel) (Gandour et al., 1987; Hewlett et al., 1997) and friction noise duration (Hewlett et al., 1997). However, there are only a few studies concerning other acoustic characteristics of EL consonants, especially the spectral peaks and F2 onset. Therefore, further analysis is necessary to understand these characteristics and how they affect EL consonant intelligibility.

Regarding the perceptual characteristics, the low intelligibility of EL consonants is mainly attributed to the voicing feature confusion for plosives, although vowel confusion also plays a role (Weiss and Basili, 1985; Weiss et al., 1979). It has been demonstrated that the VOT of EL consonants is a major cause of this confusion (Gandour et al., 1987; Hewlett et al., 1997). Further, the continuous pulsing of the EL device creates a continuously ‘voiced’ signal that causes listeners difficulty in accurately identifying voiceless consonants, causing “voiced-for-voiceless” confusion or perceptual omission of voiceless consonants. However, the above studies did not involve Mandarin EL consonants. At present, the acoustic and perceptual characteristics of Mandarin EL consonants still remain unknown.

1.2. Acoustic properties of mandarin el consonants

Due to the separation of the upper and lower respiratory tracts during laryngectomy, EL speakers are unable to access pulmonary airflow during voice and speech production. Therefore, it is expected that the smaller airflow will also lead to lower intensity, shorter consonant duration and lower spectral peaks for Mandarin EL consonants. Due to the shorter vocal tract length involved in EL speech production, the formants of EL vowels had been shown to be higher than those of normal vowels (Ng and Xiong, 2015). Therefore, it is inferred that EL consonants should have larger F2 onsets than normal consonants. On the other hand, since laryngectomy maintains the supraglottal vocal tracts intact, it is also expected that the correlation of spectral peaks and places of articulation and the correlation of F2 onsets and places of articulation should remain unchanged.

1.3. Perception of mandarin el consonants

Laryngectomy removes the larynx, leading to a lack of airflow during articulation but maintaining the major vocal tract structure for speech production. Therefore, EL consonant confusion is mainly attributed to confusion between different manners of articulation rather than confusion between different places of articulation. This has been demonstrated by previous studies about other languages. The consonant confusion in English EL speech has been mainly attributed to “voiced-for-voiceless confusion”, e. g. voiceless plosives to voiced counterparts (more than 50%), voiceless affricates to voiced counterparts (more than 35%), voiceless fricatives to voiced counterparts (more than 12%) and voiceless to sonorant (about 7%) (Weiss et al., 1985). In addition, more than 80% of voiceless aspirated stops in Thai were perceived as voiceless unaspirated stops (Gandour et al., 1987), which refers to “unaspirated-for-aspirated” confusion. VOT, low intensity and radiated noise were considered as the main factors leading to these confusions. Likewise, suffering from these acoustic defects, it is expected that the Mandarin EL consonants should also be degraded by consonant confusions between different manners of articulation, such as “voiced-for-voiceless” confusion and “unaspirated-for-aspirated” confusion.

1.4. Hypotheses

This study was designed to answer two questions concerning Mandarin EL consonants. First, what changes of acoustic properties are made in Mandarin consonants due to the use of ELs? Based on previous studies, we hypothesize that Mandarin EL consonants have smaller intensity, smaller spectral peaks, higher *F2* onset than normal consonants but unchanged correlation between acoustic properties (including spectral peak and *F2* onset) and place of articulation. Second, what confusion types are involved in Mandarin EL consonants? Our hypothesis is that Mandarin EL consonant confusion is mainly attributed to confusion between manners of articulation rather than confusion between places of articulation.

2. Method and experiments

2.1. Participants

Fifteen male laryngectomees (mean age: 69 years; range: 58 to 80 years) and fifteen male laryngeal speakers (mean age: 67 years; range: 59 to 73 years) participated in speech recordings. All subjects were Chinese natives. Nine laryngectomees spoke standard Mandarin Chinese as their primary language (originating from Peking, China) and six spoke the Wuhan dialect as their primary language (originating from Wuhan, Hubei province, China). The six Wuhan laryngectomees also spoke standard Mandarin. All laryngectomees had been using an EL as their main form of communication in daily life for a mean of approximately 3.5 years (range: 2 to 8 years) and confirmed this was their primary form of alaryngeal communication. The laryngeal speakers were recruited from the Xi’an Jiaotong University professor population and spoke standard Mandarin Chinese as their primary language. The laryngectomees and laryngeal speakers were free of any upper respiratory infection in the two weeks prior to recording. The laryngectomees reported no history of speech or language disorders other than the laryngectomy. The laryngeal speakers reported to be in good health at the time of study with no known history of voice, speech, or language disorders. For perceptual evaluation, twenty male subjects (mean: 26 years; range: 23 to 29 years) were recruited from the Xi’an Jiaotong University student population. The listeners were Chinese natives and spoke standard Mandarin Chinese as their primary language, reporting no history of speech and listening disorders. All participants volunteered for the experiments without monetary compensation.

2.2. Speech materials

- In Mandarin phonology, there are 22 consonants in total, including 21 syllable-initial consonants and 2 syllable-final consonants (⟨ng⟩ and ⟨n⟩, where ⟨n⟩ can be used as both a syllable-initial and a syllable-final consonant). Only five Mandarin consonants are voiced (⟨m, n, l, r, ng⟩), while the other 17 Mandarin consonants are voiceless. The syllable-final consonant ⟨ng⟩ is rarely confused as the other

Table 1
The Mandarin initial consonants (bold letters) and transcription symbols in International Phonetic Alphabet (enclosed in brackets).

	Labial	Alveolar	Retroflex	Alveolo-palatal	Velar
Unas-PLO	**b** [p]	**d** [t]			**g** [k]
As-PLO	**p** [p^h]	**t** [t^h]			**k** [k^h]
Unas-AFF		**z** [ts]	**zh** [ʈʂ]	**j** [tɕ]	
As-AFF		**c** [ts^h]	**ch** [ʈʂ^h]	**q** [tɕ^h]	
FRI	**f** [f]	**s** [s]	**sh** [ʂ]	**x** [ɕ]	**h** [x]
Voiced	**m** [m]	**n** [n], **l** [l]	**r** [ʐ]		

Consonants produced using the same articulation manner are listed in a same row and the consonants produced at the same articulation place are listed in the same column. Unas-PLO refers to unaspirated plosives; As-PLO are aspirated plosives; Unas-AFF are unaspirated affricates; As-AFF are aspirated affricates; FRI are fricatives and Voiced are voiced consonants.

syllable-initial consonants. Therefore, except ⟨ng⟩, only the 21 Mandarin syllable-initial consonants were considered for this study. The syllable-initial Mandarin consonants were classified by their manner and place of articulation, shown in Table 1. The speech materials were all disyllabic words formed by combing two independent monosyllables. The first syllable was fixed with the monosyllable ⟨ā⟩. The second monosyllable was formed by combining each consonant with five frequently-used rhymes. The rhymes were monophthong, diphthong, triphthong and even "vowel + consonant". The production of each token (each disyllable) was recorded in isolation. The speech materials are listed in Appendix I.

2.3. Speech material recording

For the recordings, laryngectomees and laryngeal speakers were seated in a quiet room (background noise < 40 dB, measured by a sound level meter [HT8352, HCJYET, Guangdong, China]) and asked to read the speech materials. The laryngectomees used the "Xiwang VII" (Tianchou medical machinery, Huzhou, China) device that was pre-set to a fixed fundamental frequency according to individual usage habits (mean: 56.6 Hz, ranges: 50 to 60 Hz). The specific model cannot modify its fundamental frequency after pre-setting. Each material was presented to the speakers using pinyin and a corresponding Chinese character that is frequently used in daily life. Each speaker was given a brief practice period to familiarize themselves with the speech materials, recording format, and instrumentation before the actual recording, and were instructed to read the speech materials as if communicating with a person at a distance of approximately one meter. 1575 EL samples and 1575 laryngeal samples (105 samples/subject) were recorded using a dynamic microphone (Danyin DM-099) placed 10 cm in front of the subject's mouth and digitized at 44,100 Hz and 16 bits / sample using Praat 6.0.40 (Boersma and Weenink, 2019).

2.4. Acoustic analysis

2.4.1. Speech signal pre-processing

To remove the masking effect of EL-radiated noise on consonants, the EL speech was processed using a radiated noise suppression algorithm called "multiband time-domain amplitude modulation (MTAM)" before acoustic analysis. The MTAM algorithm removes EL-radiated noise without degrading speech quality (Xiao et al., 2018). 4 clearly mispronounced tokens in laryngeal speakers were excluded for further analysis. However, potential mispronounced tokens in EL speakers were not excluded due to the poor speech quality that results in difficulty distinguishing the mispronounced tokens.

2.4.2. Time-domain characteristic analysis

Time-domain analysis was conducted based on speech intensity and consonant duration. In this study, the duration of plosives and affricates were defined as the VOT. The duration of fricatives was defined as the frication noise duration in the spectrogram. The duration of voiced consonants is the time interval between the onset of EL voice pulsing and the onset of the pulsing associated with the following vowel. For some consonants where it is hard to distinguish the boundary directly (such as ⟨l, r⟩), the intensity of formants, and in particular the first formant, was used to determine the boundary, since most voiced consonants have lower formant intensity than surrounding vowels. The intensity of consonants was defined as the average intensity (calculated using the default method of Praat) over the consonant duration, as defined above. Although the speech intensity given by Praat is not true speech intensity, it can be used to determine the relative differences between EL and normal consonants.

2.4.3. Spectral characteristic analysis

The spectral analysis of EL and normal consonants was conducted based on the spectral peak and *F2* onset. The spectral peak is mainly analyzed for affricates and fricatives, since the affricates and fricatives are characterized by a relatively stationary configuration of articulation (Maniwa et al., 2009). In this study, the spectral peak was examined using a 40 ms Hamming window placed in the middle of the frication noise. The FFT spectra were obtained using the default method of Praat.

The apparent *F2* starting frequency of a vowel preceded by an obstruent provides information about the configuration of articulation used to generate the consonant (Jongman et al., 2000). Generally, the *F2* onset rises progressively as the place of articulation moves back in the oral cavity (Alwan et al., 2011; Suchato, 2004). In addition, the first three formants (*F1, F2* and *F3*) of voiced consonants were also analyzed. In this study, a 40 ms Hamming window was placed in the middle of the consonant duration, and the formants were obtained using the default method of Praat (Burg; window duration: 20 ms; Dynamic range: 30 dB). In order to analyze the relationship of *F2* onsets with the places of articulation more clearly, syllables whose consonants were followed by a same rhyme and across at least three places of articulation were selected for F2 onset analysis, including consonants + ⟨a, ou, ong, i, e, u⟩.

Throughout this study, the mixed effect model of repeated measures analysis of variance (ANOVA) was used as the statistical analysis method. The voice source (EL and larynx) was set as between-subject factor and fixed effect, while vowels were set as the within-subject factor. The subject was set as a random effect.

2.5. Evaluation of consonant intelligibility

2.5.1. Listening task

Listeners were presented with the speech stimuli (including every recorded disyllable) through a high-fidelity sound system (EDIFIER R18T) in the same quiet room (background noise level <40 dB). The listeners were seated about 1 m away from the loudspeaker. The intensity level of the playback was adjusted by individual judges to a comfortable level. To avoid perceptual fatigue, 1575 tokens were listened over 5 days (315 tokens were tested each day). Each disyllable was played 3 times within 10 s. There was an interval of 5 s between different disyllable playbacks. Listeners had 10 minutes' break time after listening for 20 min continuously. It took about two hours to complete the listening tasks each day. The speech materials were played with a randomized order. The listeners were asked to write down what they heard, even if they thought it was meaningless.

2.5.2. Intelligibility and confusion analysis

Both the intelligibility and the confusion analysis of the EL consonants were conducted from two aspects: manners of articulation and places of articulation. The intelligibility analysis investigated strict perceptual accuracy for each consonant, and confusion analysis investigated the consonant confusion between different articulatory types (the manner of articulation or the place of articulation). Therefore, in intelligibility analysis, correct identification meant that a consonant was identified as itself, while in confusion analysis, a correct identification meant that a consonant was identified as a consonant with the same type of articulation. Then, the intelligibility and confusion results were analyzed to determine the perceptual characteristics of Mandarin EL consonants.

3. Results

3.1. Time-domain characteristics

3.1.1. Waveforms and spectrogram

Fig. 1 shows the waveforms and spectrograms of laryngeal and EL speech samples. Visually, three qualitative results can be concluded from Fig. 1. First, the amplitude of the EL speech signal is smaller than that of the laryngeal signal. Second, EL consonants have large variations across EL speakers. As shown in Fig. 1, the sound energy of /s/ produced by LS1 is much greater than that produced by ELS1 or ELS2, and some /s/ (see [a sɤ]) produced by ELS2 have extremely low levels of sound energy. Third, all EL consonant durations were injected with the continuous pulsing of the EL device, which masks the consonant signals to some extent.

3.1.2. Speech intensity

The speech intensity of EL and normal consonants is shown in Fig. 2. The speech intensity of EL consonants is in the range of 44–48 dB (mean: 46.1 dB), and that of normal consonants is in the range of 48–56 dB (mean: 52.9 dB). The speech intensity of each type of EL consonant is at least 4 dB lower than that of normal consonants. Repeated measures ANOVA with mixed effect model revealed that there is a significant difference between the intensity of each consonant type ($p < 0.01$ for each consonant type). This result indicates that the laryngectomees speaking with an EL produce significantly weaker consonant sounds than laryngeal speakers do.

3.1.3. Consonant duration

The duration of EL and normal consonants with different manners of articulation are shown in Fig. 3. For EL consonants, each type has a smaller duration than the corresponding normal ones, ($p < 0.01$ for each consonant type, except unaspirated plosives). In addition, it can be seen that the unaspirated consonants and voiced consonants have a much smaller reduction of consonant duration (average 7.6 ms and 6.6 ms, respectively) than aspirated consonants and fricatives (average 75 ms and 87 ms, respectively). These results indicate that laryngectomees produced significantly shorter consonants than laryngeal speakers speaking normally do, and EL consonants demanding higher airflow during articulation have a larger reduction of consonant duration.

3.2. Spectral characteristics

3.2.1. Spectral peak

Fig. 4 shows the spectral peaks of fricatives and affricates in EL and laryngeal speech as a function of place of articulation. For each consonant type, the EL consonants have significantly smaller spectral peaks than corresponding normal ones, ($p < 0.01$ for each consonant type). However, for both EL and normal consonants, the spectral peaks decrease as the place of articulation moves from the lip to the glottis. This result suggests that although EL consonants have smaller spectral peaks than normal consonants, they still retain the same relationship of spectral peaks and places of articulation with normal consonants.

3.2.3. Formants

Fig. 5 shows the *F2* onset of EL and laryngeal speech varying with the places of articulation. For each type, EL consonants have larger *F2* onsets than the corresponding normal consonants ($p < 0.05$ for each type, except velar + ⟨e⟩, labial + ⟨i⟩ and alveolar + ⟨i⟩). Besides, Fig. 6 shows that all the first three formants of the EL voiced consonants are also

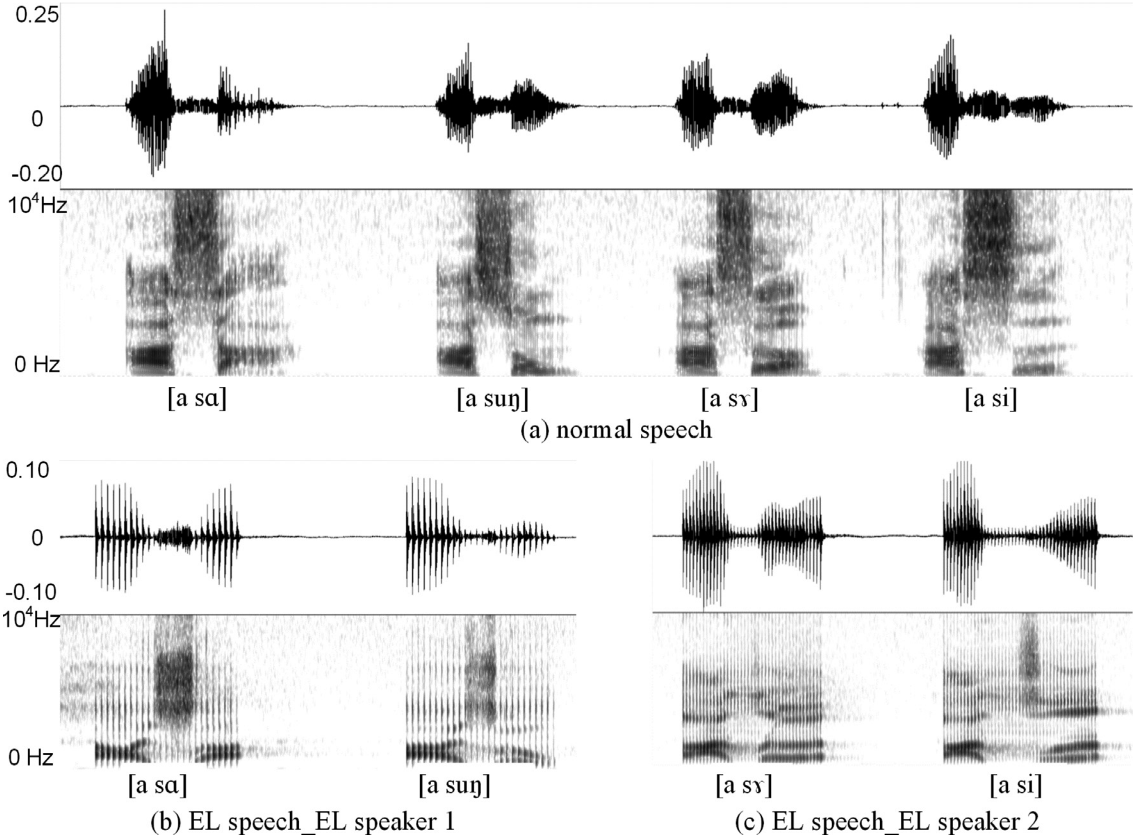

Fig. 1. Speech produced by a laryngeal speaker (LS1) and two EL speakers (ELS1 and ELS2). (a) laryngeal speech produced by LS1; (b) speech produced by ELS1; (c) speech produced by ELS2.

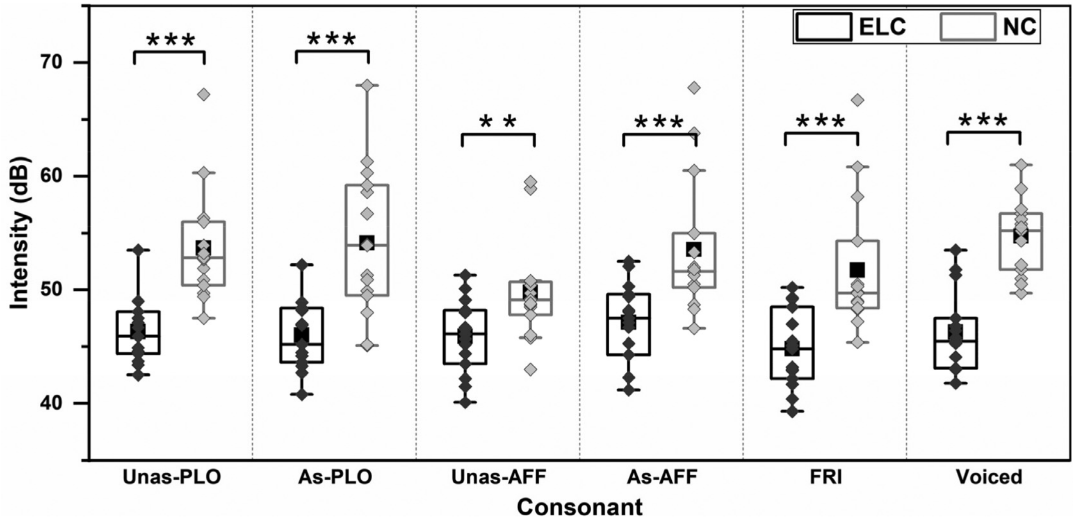

Fig. 2. Speech intensity of EL and normal consonants. ** $p < 0.01$; *** $p < 0.001$ (repeated measures ANOVA with mixed effect model). ELC is the EL consonant; NC is the normal consonant. Scatters represent individual data.

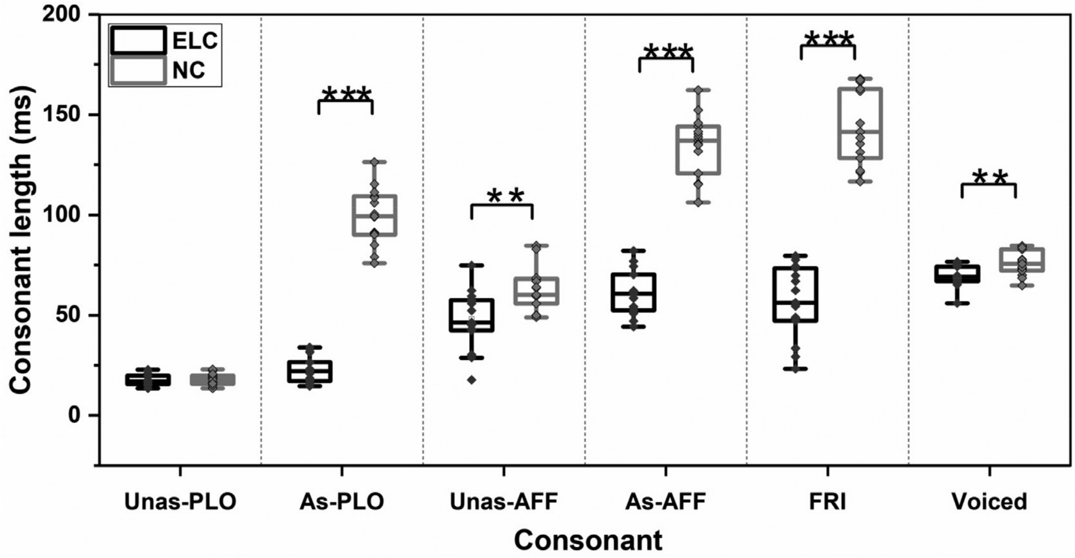

Fig. 3. Duration of EL consonants and normal consonants. ** $p < 0.01$; *** $p < 0.001$ (repeated measures ANOVA with mixed effect model). ELC represents EL consonant; NC represents normal consonant. Scatters represent individual data.

significantly larger than those of normal consonants ($p < 0.001$ for each voiced consonant). Meanwhile, Fig. 5 also shows that the *F2* onsets of EL and normal consonants vary consistently with the place of articulation moving back in the vocal tract. The *F2* onsets of both EL and normal consonants are progressively higher as the place of articulation moves back in the oral cavity, except in the velar case. These results suggest that although EL speech has higher formants (including *F2* onsets of voiceless consonants and the first three formants of voiced consonants) than laryngeal speech, they both have higher F2 onsets with the place of articulation moving to the glottis.

Based on the above acoustic analysis, the following results are derived:

(1) EL consonants have significantly lower speech intensity than normal ones;
(2) The duration of EL consonants is significantly shorter than those of normal ones, especially the aspirated consonants and fricatives;
(3) EL consonants have lower spectral peaks, but the relationship between their spectral peaks and place of articulation is similar to that of normal consonants.
(4) The EL consonants have significantly higher formants than normal consonants do, however, the relationship between *F2* onset and place of articulation remains unchanged.

3.3. Perceptual characteristics

3.3.1. Intelligibility

The intelligibility of EL consonants as a function of the place and manner of articulation are respectively presented in Figs. 7 and 8. The results shown in both figures indicate that most EL consonants have low intelligibility. Fig. 7 shows that the intelligibility of EL consonants has no apparent correlation with place of articulation moving from lip to glottis (59% on average, range: 48%–69%). In contrast, the intelligibility of EL consonants is closely related to the manner of articulation, as shown in Fig. 8. The intelligibility of voiced consonants is highest (86.3%) and the intelligibility of unaspirated consonants is much larger (69.5%) than that of aspirated consonants and fricatives (41.2% and 44.9% respectively). These results indicate that the intelligibility of EL consonants is more closely related with the manner of articulation rather than the place of articulation. Generally, consonants demanding

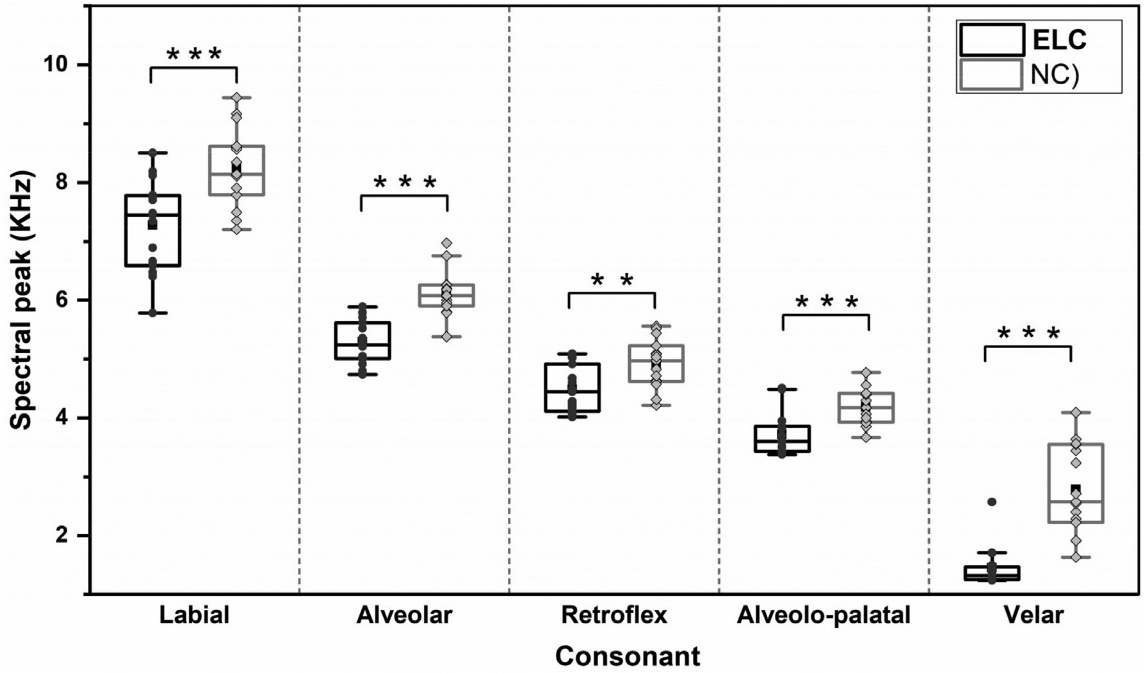

Fig. 4. Spectral peaks of fricatives and affricates in EL and laryngeal speech. ** $p < 0.01$; *** $p < 0.001$ (repeated measures ANOVA with mixed effect model). ELC represents EL consonant; NC represents normal consonants. Scatters represent individual data.

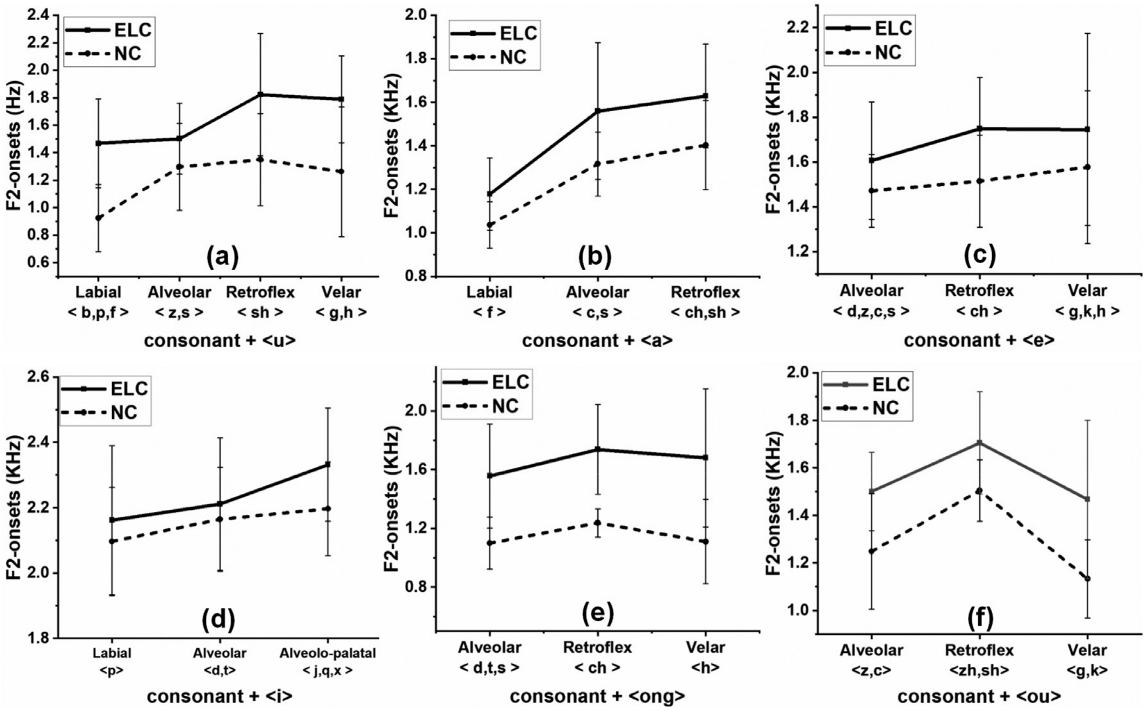

Fig. 5. *F2* onset frequencies of EL and normal consonants as a function of place of articulation. ELC are EL consonants; NC are normal consonants.

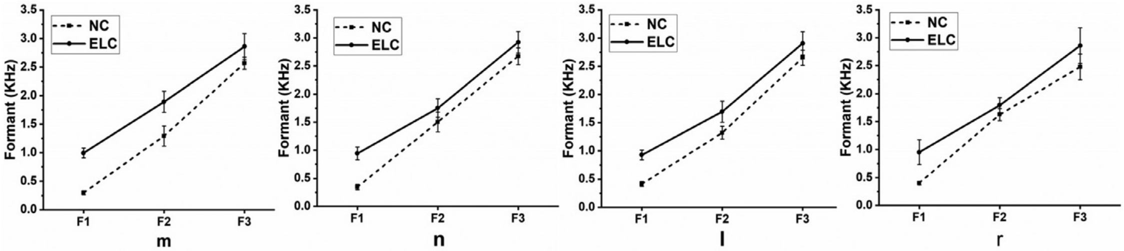

Fig. 6. First three formants of voiced consonants in EL and laryngeal speech. ELC are EL consonants; NC are normal consonants.

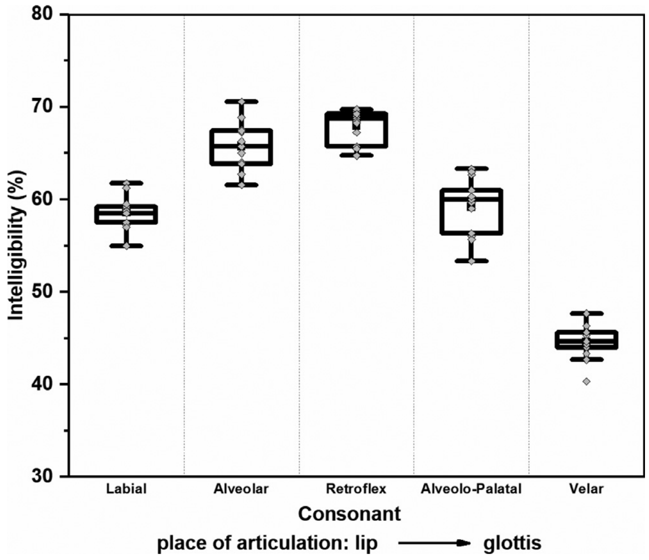

Fig. 7. Variation of EL consonants' intelligibility with place of articulation moving from lip to glottis. Scatters represent individual data.

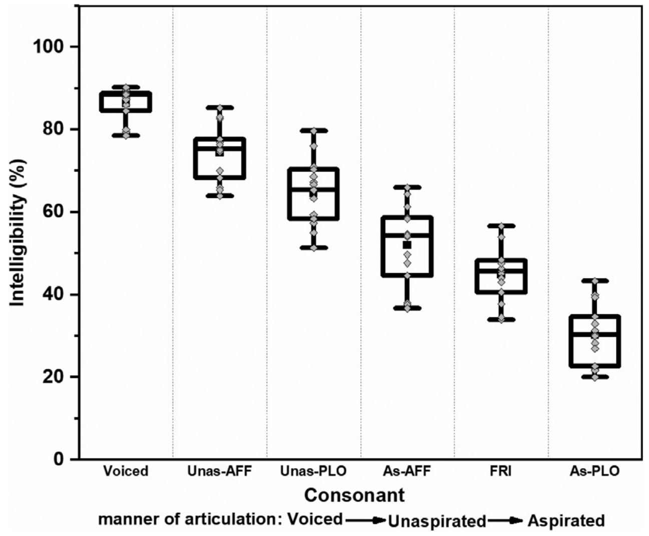

Fig. 8. Variation of EL consonants' intelligibility with manner of articulation as airflow demand increases. Scatters represent individual data.

less airflow during articulation have larger intelligibility in Mandarin EL speech.

3.3.2. Consonant confusion

We used confusion matrices between different places and manners of articulation to determine the Mandarin EL consonant confusion patterns. The confusion matrix produced by different manners of articulation is shown in Table 2. According to this result, consonant confusion exists between EL consonants with different manners of articulation, and three patterns are mainly responsible for the confusion between different manners of articulation. (1) The EL aspirated consonants are easily mistaken as their unaspirated counterparts (33.2% as aspirated plosives and 26.3% as aspirated affricates), which in this study is referred to as "unaspirated-for-aspirated"; (2) Voiceless consonants are easily identified as voiced consonants, which is referred to as "voiced-for-voiceless" confusion; (3) Voiceless consonants are easily missed, which is referred to as "perceptual omission". The fricatives and aspirated consonants suffer larger "perceptual omission" (33.0% and 16.1%, respectively) than unaspirated consonants (10.4%) do.

Table 2
Confusion matrix of EL consonants by different manners of articulation (%).

	Unas-PLO	As-PLO	Unas-AFF	As-AFF	FRI	Voiced	omission
Unas-PLO	67.0	4.3	0.6	0.0	0.3	17.3	10.5
As-PLO	33.2	31.1	0.4	0.0	2.2	14.3	18.8
Unas-AFF	1.5	0.0	77.6	0.7	0.7	9.4	10.2
As-AFF	0.2	0.0	26.3	53.7	0.3	6.2	13.4
FRI	1.1	0.0	11.2	0.6	46.9	7.1	33.0
Voiced	3.5	0.0	0.6	0.0	0.1	90.3	5.4

The stimulus phonemes are listed vertically, while perceived phonemes are indicated horizontally.

Table 3
Confusion matrix of EL consonants by different places of articulation (%).

	Labial	Alveolar	Retroflex	Alveolo-palatal	Velar	omission
Labial	79.7	3.6	0.8	0	0.9	15.1
Alveolar	0.45	85.2	3.4	0.04	0.2	10.8
Retroflex	0.02	3.6	84.7	0	0.5	11.2
Alveolo-palatal	0.6	4.4	3.1	75.2	0.1	16.6
Velar	0.7	4.7	1.8	0.2	60.3	32.3

Stimulus phonemes are listed vertically while perceived phonemes are indicated horizontally.

The confusion matrix of EL consonants with different places of articulation is shown in Table 3. "Perceptual-omission" is still prominent (average 17.2%, range: 10.8%–32.3%) while other confusion types are not pronounced (smaller than 7% on average for each consonant type). This suggests that confusion due to the different places of articulation contributes very little to the poor intelligibility of EL speech. Further, it can be inferred that the "voiced-for-voiceless" and "unaspirated-for-aspirated" confusions were mainly due to confusions produced at the same place of articulation.

Based on the perceptual test analysis, the following conclusions are drawn:

(1) Consonant confusion in EL speech is mainly due to voiceless consonants, leading to the low intelligibility of voiceless consonants in Mandarin EL speech;
(2) Confusion of voiceless consonants mainly occurs due to different manners rather than different places of articulation in Mandarin EL speech;
(3) Consonant confusion between different manners of articulation in Mandarin EL speech mainly manifests through three patterns: "unaspirated-for-aspirated" confusion, "voiced-for-voiceless" confusion and "perceptual-omission" confusion.

4. Discussion

4.1. Abnormal acoustic characteristics

The speech intensity analysis indicates that laryngectomees speaking with an EL produce significantly weaker consonants than laryngeal speakers normally do. This is expected, as due to removal of the larynx, laryngectomees cannot produce laryngeal voice using pulmonary airflow. Due to this inability, the consonants produced by laryngectomees using an EL do not achieve the intensity of normally produced consonants. In addition, the vocal efficiency of EL speech production is also far smaller than the vocal efficiency of laryngeal speech production (Wu et al., 2017), therefore, the intensity of EL speech, including voiced consonants and vowels, is much lower than laryngeal speech.

Also suffering from the inability of utilizing pulmonary airflow during articulation, the duration of consonants demanding higher airflow is reduced more. Shortened VOTs are also observed for word-initial voiceless stop consonants in English, which are mostly reduced to less than 30 ms (Hewlett et al., 1997; Weiss et al., 1979) and are even reduced to lower than 20 ms for Thai EL speech (Gandour et al., 1987). For EL speakers, only a very limited volume of buccal air can be used to produce consonants; this volume can only partly compensate the airflow demand for producing unaspirated consonants, but this is far from the airflow demand for producing aspirated consonants and fricatives. Therefore, the durations of aspirated consonants and fricatives are significantly shortened, closely approaching their unaspirated counterparts.

The spectral analysis reveals that EL consonants have lower spectral peaks than normal ones, maintaining however the correlation of spectral peaks and places of articulation. The spectral peak is proportional to the airflow velocity and inversely proportional to the diameter of the constriction (Narayanan and Alwan, 2000). The laryngectomy removes the larynx of patients, but maintains the other articulation structures (such as oral cavity and tongue) intact. Under the condition of unchanged constriction, the spectral peaks of EL consonants inevitably become smaller due to the lack of airflow. Due to the remaining of the major structures of articulation, the information of associated with the place of articulation in the spectral peaks remains present.

Considering the formants, EL speech has significantly larger formants than laryngeal speech (including *F2* onset and the first three formants of voiced consonants). Except for EL speech, higher formant frequencies are found in esophageal and tracheoesophageal speech (Liao, 2016; Ng and Xiong, 2015). Generally, a shorter vocal tract leads to higher formant frequencies. Although the vocal tract configuration (length and volume) in laryngectomized individuals has been shown not to be significantly different than that of laryngeal speakers (Ng et al., 2018), the voice source for using the EL is located somewhere beyond the glottis rather than the end of the vocal tract, which is the normal laryngeal voice location. This shortens the prior cavity involved in articulation, which is the cause for the higher formants of EL consonants.

As analyzed above, the *F2* onset frequencies in EL speech are also higher than those in laryngeal speech, due to the shortened prior cavity involved in articulation. However, it is interesting that the correlation of the *F2* onset and place of articulation in EL speech is consistent with that of laryngeal speech. It has been demonstrated that the apparent *F2* starting frequency of a vowel preceded by an obstruent provides information about the articulatory configuration used to generate the consonant (Sussman et al., 1991), since the articulatory configuration of vowel onset is close to that of the preceding consonant (Suchato, 2004). Laryngectomy removes the larynx, but preserves the supraglottal vocal tract structure intact, so laryngectomees can shape any articulatory configuration. Therefore, similarly to the spectral peak, the unchanged correlation between the F2 onset and place of articulation is mainly due to the intact supraglottal vocal tract structure.

4.2. Acoustic mechanisms for consonant confusion

At present, there is no evidence showing that low speech intensity is related to any specific consonant confusion pattern. However, it is expected that listeners have difficulty in perceiving the speech information correctly at low speech intensity, resulting in some consonant confusions. For EL speech in particular, the continuous pulsing of the EL device will increase the level of noise in EL speech, masking the EL consonants and leading to “voiced-for-voiceless” confusions. The results demonstrate that, generally, larger intensity reduction corresponds to lower intelligibility for EL voiceless consonants. Previous studies have shown that increasing the intensity of consonants can effectively improve speech intelligibility at low speech intensity (Digiovanni and Stover, 2008; Jayan and Pandey, 2015; Sarath and Jayan, 2017). Therefore, the low speech intensity of EL consonants is partly responsible for the low intelligibility of EL consonants.

The “unaspirated-for-aspirated” confusion in Mandarin EL consonants is mainly attributed to the shortened VOT of aspirated consonants. VOT is considered as a key factor for distinguishing aspirated from unaspirated consonants (Liu, 2011; Wong, 2007). In general, aspirated consonants have longer VOTs than their unaspirated counterparts (Qi and Zhang, 1982). The VOT of Mandarin EL aspirated consonants is reduced and approaches the VOT of unaspirated counterparts (see Fig. 3), which results in “unaspirated-for-aspirated” confusions (33% for aspirated plosive and 26.3% for aspirated affricates, see Table 2). Therefore, VOT reduction of aspirated consonants is mainly responsible for the “unaspirated-for-aspirated” confusion in Mandarin EL speech.

Both the “voiced-for-voiceless” confusion and “perceptual omission” are mainly due to low intensity or omission of consonants and continuous EL pulsing in Mandarin EL speech. ELs produce continuous a pulsing of the voice source throughout the speaking duration. The EL voice source is a periodic vibration signal that resembles the source of the vowels and voiced consonants. During a consonant, the EL voice source simultaneously reconstructs a voiced speech signal and in the process produces radiated noise that masks weak consonants to some extent. When the speech intensity of EL consonants is too low to perceive, listeners will identify this interval as a voiced interval. If the consonant has the same place of articulation with a voiced consonant, this duration is easily identified as a voiced consonant, i.e. a “voiced-for-voiceless” confusion occurs. If the consonant does not have the same place of articulation with any other voiced consonant, this duration is easily identified as a vowel or a semivowel, i.e. an “omission”.

Consonant confusion among Mandarin EL consonants with different place of articulation is not prominent. Spectral peaks and *F2* onsets are closely related to the place of articulation (Alwan et al., 2011; Jongman et al., 2000). It was shown that, due to the intact supraglottal vocal tract, the information of place of articulation is maintained in the spectral peaks and *F2* onset frequencies: as the place of articulation moves from lip to glottis, the spectral peak decreases and *F2* onset increases (see Fig. 4 and Fig. 6). Therefore, consonant confusions rarely occur among different places of articulation.

In summary, confusions in voiceless consonants are the major causes of low intelligibility of Mandarin EL consonants, and can be attributed to three patterns: (1) the “unaspirated-for-aspirated” confusion caused by shortened VOT of aspirated consonants; (2) the “voiced-for-voiceless” confusion caused by low consonant-intensity and strong radiated noise of EL; and (3) the perceptual omission caused by low consonant-intensity or consonant-omission. The two research hypotheses are therefore supported by these results. It is difficult to resolve consonant confusions by improving the EL device with advanced materials or construction. In recent years, many signal-processing methods have been successfully utilized to enhance EL speech, such as voice-conversion technology (Doi et al., 2014; Nakamura et al., 2012). Since EL speech involves the vocal tract characteristics, the vocal tract transfer function of EL consonant duration can be extracted and utilized to resynthesize a consonant duration that can improve the intensity of EL consonants effectively or extend the VOT of aspirated consonants. Based on this assumption, the next work is to improve the intelligibility of Mandarin EL speech by extending the VOT of aspirated consonants and compensating for low consonant intensity or consonant omission. In addition, enhancing low-frequency deficits, reducing EL device noise, and greater frequency variation may lead to improvements in Mandarin EL speech quality and / or intelligibility (Meltzner and Hillman, 2005).

5. Conclusion

This study examined the acoustic and perceptual characteristics of Mandarin consonants produced by EL users and laryngeal speakers. Acoustic analysis revealed that: (1) EL consonants had significantly lower intensity than normal consonants; (2) EL consonants had significantly shorter durations than normal consonants; (3) EL consonants had lower spectral peaks and higher *F2* onsets than normal consonants, but both quantities maintained their correlation with the place of articulation. Perceptual results revealed that consonant confusion mainly occurs among voiceless consonants with different manners of articulation rather than with different places of articulation. Consonant confusion among different manners of articulation is attributed to three patterns: (1) the "unaspirated-for-aspirated" confusion caused by the shortened VOT of aspirated consonants; (2) the "voiced-for-voiceless" confusion caused by low intensity and continuous EL pulsing; and (3) "perceptual omission" caused by low intensity or omission of EL consonants. Hence, raising the intensity of EL consonants, extending the VOT of aspirated consonants or reconstructing missed EL consonants are beneficial and promising ways to improve intelligibility of EL consonants in Mandarin speech in the future.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Ke Xiao: Conceptualization, Formal analysis, Investigation, Methodology, Writing - review & editing. **Bo Zhang:** Formal analysis, Investigation. **Supin Wang:** Project administration. **Mingxi Wan:** Funding acquisition, Project administration. **Liang Wu:** Funding acquisition, Supervision, Writing - review & editing.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Nos. 81771854, 11274250, and 11874049).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.specom.2020.06.004.

References

Alwan, A., Jiang, J., Chen, W., 2011. Perception of Place of Articulation for Plosives and Fricatives in Noise. Speech Commun 53, 195–209.

Basha, S.K., Pandey, P.C., 2012. Real-time enhancement of electrolaryngeal speech by spectral subtraction. Communications 1–5.

Boersma, P., Weenink, D., 2019. Praat: doing phonetics by computer, http://www.fon.hum.uva.nl/praat/ (version number: 6040).

Chen, W., Zheng, R., Baade, P.D., Zhang, S., Zeng, H., Bray, F., Jemal, A., Yu, X.Q., He, J., 2016. Cancer statistics in China, 2015. Ca Cancer J Clin 66, 115–132.

Digiovanni, J.J., Stover, A.K., 2008. The role of consonant duration and amplitude processing on speech intelligibility in noise. Journal of the Acoustical Society of America 123, 3865.

Doi, H., Toda, T., Nakamura, K., Saruwatari, H., Shikano, K., 2014. Alaryngeal Speech Enhancement Based on One-to-Many Eigenvoice Conversion. IEEE/ACM Transactions on Audio Speech & Language Processing 22, 172–183.

Espy-Wilson, C.Y., Chari, V.R., MacAuslan, J.M., Huang, C.B., Walsh, M.J., 1998. Enhancement of electrolaryngeal speech by adaptive filtering. Journal of Speech, Language, and Hearing Research. 41, 1253–1264.

Faulkner, A., Rosen, S., 1999. Contributions of temporal encodings of voicing, voicelessness, fundamental frequency, and amplitude variation to audio-visual and auditory speech perception. Journal of the Acoustical Society of America. 106, 2063.

Gandour, J., Weinberg, B., Petty, S.H., Dardarananda, R., 1987. Voice onset time in Thai alaryngeal speech. Journal of Speech & Hearing Disorders 52, 288–294.

Gandour, J., Weinberg, B., Petty, S.H., Dardarananda, R., 1988. Tone in Thai alaryngeal speech. Journal of Speech and Hearing Disorders 53, 23–29.

Hewlett, N., Cohen, W., Macintyre, C., 1997. Perception and production of voiceless plosives in electronic larynx speech. Clin Linguist Phon 11, 1–22.

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.

Jayan, A.R., Pandey, P.C., 2015. Automated modification of consonant–vowel ratio of stops for improving speech intelligibility. Int J Speech Technol 18, 113–130.

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

Kaye, R., Tang, C.G., Sinclair, C.F., 2017. The electrolarynx: voice restoration after total laryngectomy. Medical Devices 10, 133–140.

Liao, J.S., 2016. An Acoustic Study of Vowels Produced by Alaryngeal Speakers in Taiwan. Am J Speech Lang Pathol 25, 481–492.

Liu, H., Ng, M.L., 2007. Electrolarynx in voice rehabilitation. Auris Nasus Larynx 34, 327–332.

Liu, H., Wan, M., Ng, M.L., Wang, S., Lu, C., 2006. Tonal Perceptions in Normal Laryngeal, Esophageal, and Electrolaryngeal Speech of Mandarin. Folia Phoniatrica Et Logopaedica Official Organ of the International Association of Logopedics & Phoniatrics 58, 340–352.

Liu, J.T., 2011. An Overview of Voice Onset Time. Overseas English 10, 335–336.

Maniwa, K., Jongman, A., Wade, T., 2009. Acoustic characteristics of clearly spoken English fricatives. Journal of the Acoustical Society of America 125, 3962–3973.

Meltzner, G.S., Hillman, R.E., 2005. Impact of aberrant acoustic properties on the perception of sound quality in electrolarynx speech. Journal of Speech Language & Hearing Research Jslhr 48, 766–779.

Meltzner, G.S., Hillman, R.E., Stevens, K.N., 2001. Analysis of acoustic factors contributing to poor quality of electrolaryngeal speech. Journal of the Acoustical Society of America 110, 2764.

Nagle, K.F., Eadie, T.L., Wright, D.R., Sumida, Y.A., 2012. Effect of fundamental frequency on judgments of electrolaryngeal speech. Am J Speech Lang Pathol 21, 154–166.

Nakamura, K., Toda, T., Saruwatari, H., Shikano, K., 2012. Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech. Speech Commun 54, 134–146.

Narayanan, S., Alwan, A., 2000. Noise source models for fricative consonants. IEEE Transactions on Speech & Audio Processing 8, 328–344.

Ng, M.L., Lerman, J., Gilbert, H., 1998. Perceptions of tonal changes in normal laryngeal, esophageal, and artificial laryngeal male Cantonese speakers. Folia phoniatrica et logopaedica 50, 64–70.

Ng, M.L., Xiong, M., 2015. Chinese Alaryngeal Speech Rehabilitation and Their Acoustical Characteristics: a Comprehensive Review. Rehabilitation Medicine 25, 44–49.

Ng, M.L., Yan, N., Chan, V., Chen, Y., Lam, P., 2018. A Volumetric Analysis of the Vocal Tract Associated with Laryngectomees Using Acoustic Reflection Technology. Folia Phoniatrica et Logopaedica. 44–49.

Niu, H.J., Wan, M.X., Wang, S.P., Liu, H.J., 2003. Enhancement of electrolarynx speech using adaptive noise cancelling based on independent component analysis. Med Biol Eng Comput 41, 670–678.

Owren, M.J., Cardillo, G.C., 2006. The relative roles of vowels and consonants in discriminating talker identity versus word meaninga. Journal of the Acoustical Society of America 119, 1727–1739.

Pandey, P.C., Basha, S.K., 2010. Enhancement of electrolaryngeal speech by spectral subtraction, spectral compensation, and introduction of jitter and shimmer. In: Proc. 20th International Congress on Acoustics (ICA 2010), pp. 23–27.

Qi, S., Zhang, J., 1982. A study of duration of Chinese consonants. Acta Acustica 7, 8–13.

Qi, Y.Y., Weinberg, B., 1991. Low-Frequency Energy Deficit in Electrolaryngeal Speech. Journal of Speech & Hearing Research 34, 1250–1256.

Saikachi, Y., Stevens, K.N., Hillman, R.E., 2009. Development and Perceptual Evaluation of Amplitude-Based F0 Control in Electrolarynx Speech. Journal of Speech Language & Hearing Research 52, 1360–1369.

Sarath, P.G., Jayan, A.R., 2017. Speech intelligibility enhancement on android platform by consonant-vowel-ratio modification. In: International Conference on Next Generation Intelligent Systems, pp. 1–5.

Shigeki, M., Toshihiro, K., Yu, S., 2006. Voiceless affricate/fricative distinction by frication duration and amplitude rise slope. Journal of the Acoustical Society of America. 120, 1600.

Sluis, K.E.V., Molen, L.V.D., Son, R.J.J.H.V., Hilgers, F.J.M., Bhairosing, P.A., Brekel, M.W.M.V.D., 2018. Objective and subjective voice outcomes after total laryngectomy: a systematic review. European Archives of Oto-Rhino-Laryngology 275, 11–26.

Suchato, A., 2004. Classification of stop consonant place of articulation [D]. Department of Electrical Engineering and Computer Science. Massachusetts Institute of Technology.

Sussman, H.M., McCaffrey, H.A., Matthews, S.A., 1991. An investigation of locus equations as a source of relational invariance for stop place categorization. J. Acoust. Soc. Am. 90, 1309–1325.

Verkerke, G.J., Thomson, S.L., 2014. Sound-producing voice prostheses: 150 years of research. Annu Rev Biomed Eng 16, 215–245.

Wan, C., Wang, E., Wu, L., Wang, S., Wan, M., 2012. Design and evaluation of an electrolarynx with tonal control function for Mandarin. Folia Phoniatrica Et Logopaedica 64, 290–296.

Wong, C.J., 2007. Voice onset time (VOT) characteristics of esophageal, tracheoesophageal and laryngeal speech of Cantonese. Cantonese Dialects 52 (3), 780–789.

Wang, L., Qian, Z., Feng, Y., Niu, H., 2018. Design and Preliminary Evaluation of Electrolarynx with F0 Control Based on Capacitive Touch Technology. IEEE Transactions on Neural Systems & Rehabilitation Engineering A Publication of the IEEE Engineering in Medicine & Biology Society 26, 629.

Weiss, M.S., Basili, A.C., 1985. Electrolaryngeal Speech Produced by Laryngectomized SubjectsPerceptual Characteristics. Journal of Speech, Language, and Hearing Research. 28, 294–300.

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.

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 Transactions on Biomedical Engineering 60, 1965–1974.

Wu, L., Xiao, K., Supin, W., Mingxi, W., 2017. Vocal efficiency of electrolaryngeal speech production. Speech Commun 89, 17–24.

Xiao, K., Wang, S., Wan, M., Wu, L., 2018. Radiated Noise Suppression for Electrolarynx Speech Based on Multiband Time-Domain Amplitude Modulation. In: IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26, pp. 1585–1593.