Reconstruction of Mandarin Electrolaryngeal Fricatives With Hybrid Noise Source

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

***Abstract*—The Mandarin electrolaryngeal (EL) speech is suffering from severe fricative confusion due to improper EL source in EL speech production and abnormal physiological structure of vocal tract in the laryngectomized condition. To reduce the fricative confusions, this paper proposes a hybrid noise source by combining the typical natural fricative sources and compensation sources that consider the acoustic defects in the frequency domain caused by the truncated vocal tract and abnormal source location in EL speech production. All parameters of the model are fricative-specific and the parameters of the compensation sources are determined by analyzing the vocal tract transfer functions before and after the laryngectomy. All five Mandarin fricatives are produced by laryngectomized subjects with an experimental EL system loading the hybrid noise source and the wideband noise source. The acoustic and perceptual features of these reconstructed EL fricatives are analyzed and evaluated by comparing with the conventional EL fricatives and normal fricatives. The results indicate that the hybrid noise source successfully improves the acoustic properties of the EL fricatives by forming better spectral shapes, raising the frequencies of average energy concentration, and producing better spectral skewness and kurtosis. Finally, due to these improvements of acoustic properties, the hybrid noise sources achieve much larger intelligibility for EL fricatives than the wideband noise source and the conventional EL source. Thus, the hybrid noise source is an effective, feasible, and promising method of reducing the severe fricative confusions and improving the intelligibility of EL speech.**

***Index Terms*—Acoustic analysis, electrolarynx, fricative reconstruction, intelligibility, voice source.**

I. Introduction

LARYNGECTOMY is an effective and widely used surgical treatment for laryngeal cancers, especially in China where over 25,000 patients are diagnosed with the laryngeal cancer each year [1], and most of them have to undergo the laryngectomy for survival. However, the surgery will lead to the inability of producing natural voice due to the removal of vocal folds. There are mainly three speech rehabilitation options for laryngectomees: esophageal speech, trachea-esophageal speech and EL speech. Each method has their own advantages and disadvantages [2]. Both esophageal speech and trachea-esophageal speech have better speech quality and intelligibility than EL speech [2], [3]. However, esophageal speech is too hard to learn for many laryngectomees and the trachea-esophageal speech is limited by the prosthetic devices that are easily blocked by mucus, causing complications [3], [4]. Researches indicated that, due to the advantages of easy learning, easy operation and continuous output, more than 50% of laryngectomees rely on EL as their major communication method [5]. However, the poor intelligibility and acceptability of EL speech still limit the further application of the commercial electrolarynx [6], [7].

Many researches have investigated the poor intelligibility of EL speech. The severe EL consonant confusion is mainly responsible for the poor intelligibility of EL speech, although the EL vowel confusion, the lack of fundamental frequency variation, and the leakage of self-generated EL noise also make the EL speech perceived as unclear mechanic and robotic sounds [6], [8], [9]. Aiming at these shortcomings, Qi [10] improved the intelligibility of EL speech by compensating the low-frequency energy deficits. Saikachi [11], Wan [12] and Wang [13] also increased the EL speech intelligibility by realizing the control of fundamental frequency. Basha *et al.* [14] and Niu *et al.* [15] enhanced the quality and clearness of EL speech through reducing the leakage noise and environmental noise. However, after these efforts, the EL speech intelligibility is still unsatisfying, since few attentions are paid to reducing the consonant confusion in EL speech. Therefore, it is necessary and imperative to improve the EL consonant quality and reduce the perceptual confusion for improving the EL speech intelligibility.

In our previous study, we investigated acoustical and perceptual characteristics of Mandarin consonants in EL speech [16]. It was found that the fricative consonant is frequently perceived as voiced consonants and unaspirated consonants, contributing much to the severe consonant confusion and the low intelligibility of EL speech. Furthermore, the conventional EL source in current commercial devices is shown improper for EL consonant production and is a key factor responsible for the perception confusion [16], [17]. In Mandarin phonology, all the Mandarin fricatives are voiceless consonants, which are generally produced by turbulent noise sources formed by the pulmonary airflows breaking through the vocal tract constriction or striking the vocal tract wall [18]. However, the surgical removal of the larynx makes the laryngectomees difficult in producing a noise source by using the pulmonary airflows, and the conventional EL source is always a periodic vibration

Manuscript received June 26, 2018; revised September 21, 2018 and November 1, 2018; accepted November 6, 2018. Date of publication November 9, 2018; date of current version November 29, 2018. This work was supported by the National Natural Science Foundation of China under Grants 11274250, 81771854, and 11404256. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Tan Lee. *(Corresponding author: Liang Wu.)*

The authors are with the Key Laboratory of Biomedical Information Engineering of Ministry of Education and Department of Biomedical Engineering, School of Life Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: xjtuxk@stu.xjtu.edu.cn; spwang@mail.xjtu.edu.cn; mxwan@mail.xjtu.edu.cn; liangwu@xjtu.edu.cn).

Digital Object Identifier 10.1109/TASLP.2018.2880607

2329-9290 © 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

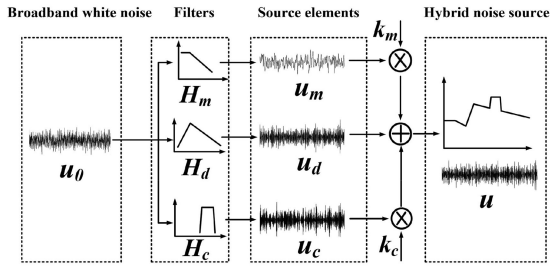

Fig. 1. The synthetic procedure of the hybrid noise source. The u_0 is the wideband white noise, and the *u* is the hybrid noise source. The u_m, u_d, and u_c are monopole source, dipole source, and compensation source respectively. The H_m, H_d, and H_c are the filters used for producing the monopole source, dipole source, and compensation source. The k_m and k_c are the scaling factors of the monopole source and the compensation source.

signal but not a noise source. Therefore, the conventional periodic voice source will certainly produce confused fricatives sounding like voiced phonemes, indicating that developing an appropriate voice source for fricative consonant production should be a direct way to improve the quality of EL fricatives and the intelligibility of EL speech.

Until now, a few studies have made efforts to ameliorate EL voice source according to the specific process of EL speech reconstruction. Particularly, Wu *et al.* [19] effectively improved the vowel intelligibility by using a vowel-specific periodic voice source with compensation of vocal tract characteristics, indicating that source-filter theory is working in EL speech production and improvement of EL voice source is able to improve the intelligibility of EL speech. In speech production models, a wideband noise is always utilized as the excitation source for voiceless consonants [20]. However, the abnormal vocal tract transfer functions of laryngectomy caused by truncated vocal tracts suggest that the wideband noise source is not the best choice for reconstructing EL fricatives. Thus, a fricative-specific voice source is necessary to reduce the perceptual confusion of different EL fricatives.

In this work, a hybrid noise source was proposed to improve the acoustical quality and the intelligibility of the Mandarin fricatives in EL speech. This source considers the acoustic characteristics of different fricatives and the vocal tract characteristics of laryngectomized subjects, aiming to increase discrimination of the fricatives and reduce the perception confusions in EL speech. The details of the source model and the experimental setup for EL speech reconstruction are described in Sec. II and III. Then, the EL fricatives produced by the hybrid noise source were evaluated from aspects of acoustical and perceptual features in Sec. IV. Finally, the feasibility and effectiveness of the hybrid noise source are discussed through a comparison with the EL fricatives produced by the wideband noise source and conventional EL source in Sec. V.

II. Hybrid Noise Source Model

As shown in Fig. 1, the hybrid noise source is a time domain waveform but modeled in the frequency domain by combining spectral features of typical fricative sources and compensation sources. The typical fricative sources were widely used noise sources in speech synthesis, which are able to generate essential spectral characteristics of the turbulent flow-induced source for fricative consonants in normal production. The compensation sources were intended to compensate the energy deficits caused by truncated vocal tract after the surgery and abnormal voice source location in EL speech production. Finally, all the typical fricative sources and the compensation sources determined the spectral features of the hybrid noise source.

TABLE I
The Synthesis Parameters of the Dipole Sources

fricatives	F_{peak}(Hz)	LPF order	HPF order
x	5000	2	4
h	5000	2	1
s	5000	2	4
sh	5000	2	4
f	5000	2	1

The LPF and HPF represent the low pass filter high pass filter, respectively.

A. Typical Fricative Sources

In this study, the typical fricative source consisted of two canonical sources, i.e., a monopole source and a dipole source as shown in Fig. 1 [18]. The spectrum of the monopole source was shaped like a low-pass filter with a fixed cutoff frequency *Fc* as in previous works [18], [21]. The spectral shape of the dipole source was a broad peak with the peak frequency F_{peak} and the spectral attenuation slopes T_{LF} and T_{HF}. In this model, the dipole source was synthesized by a low-pass filter and a high-pass filter with the same cutoff frequency F_{peak}, and the T_{LF} and T_{HF} were controlled by the low-pass filter order and high-pass filter order, respectively.

Since the effect of the monopole source is much less than the dipole source, the monopole source spectrum in this study was defined as a second-order low-pass filter with a fixed cutoff frequency of 1000 Hz for all the fricatives [18], [21]. For the dipole source, Narayanan [18], [22] found that the peak frequency was in the range of 3–6 kHz, and the slopes T_{LF} and T_{HF} were in the range of 6–24 dB/oct and -14–0 dB/oct respectively. For simplicity, the frequency peak F_{peak} was set as 5 kHz for all fricatives, while the filter orders were set according to different fricatives as listed in Table I.

B. Compensation Source

In the EL speech production, the truncated vocal tract after the laryngectomy and the abnormal source location where the EL is placed always cause significant energy attenuation in certain frequency ranges [8]. Therefore, the compensation source was designed as a band-pass filter with two pass-band edge frequencies F_L and F_H and one filter-order parameter T_C.

To determine the parameters, the vocal tract transfer functions (VTTFs) of laryngectomees before and after the surgery were estimated and compared to find the regions of the energy attenuation. Considering the same situation as in Wu's work that it is impossible to obtain the intact vocal tract of the laryngectomees [19], vocal tract area functions measured from the magnetic

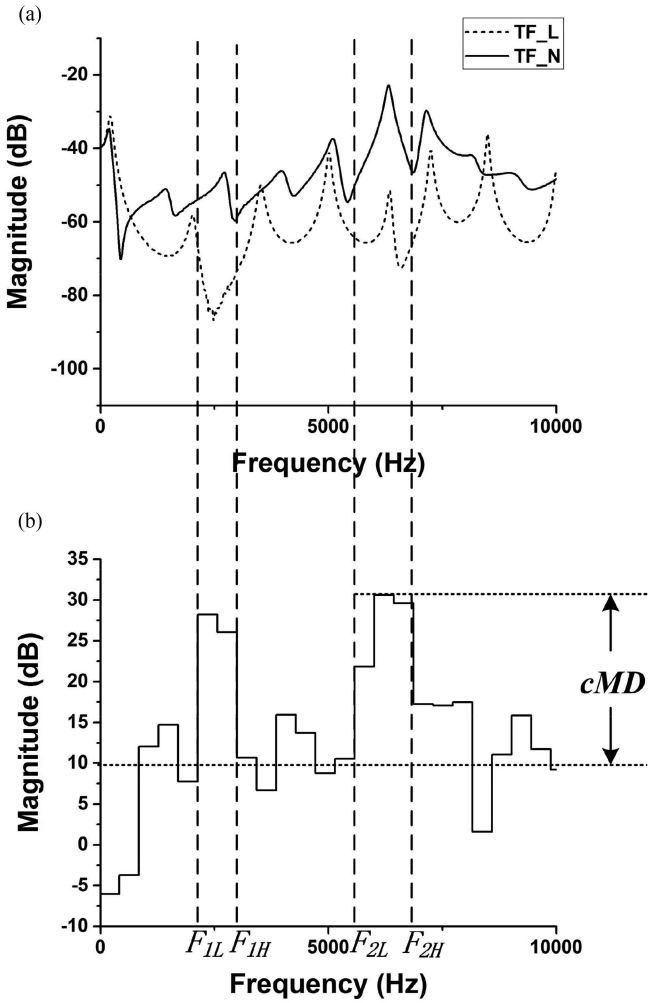

Fig. 2. (a) The vocal tract transfer functions of the fricative /s/ in the conditions of normal speech production (TF_N) and EL speech production (TF_L). (b) The segmental magnitude difference between the TF_N and TF_L. It is formed by the average magnitude differences of each segment (about 400 Hz/segment). The dash lines mark out two energy defects regions [F_{1L}, F_{1H}] and [F_{2L}, F_{2H}] in the TF_L. The dot lines mark out the magnitude difference that is expected to be compensated in the energy attenuation region (*cMD.*). The *cMD* is ensured by the largest magnitude difference minus the average magnitude difference.

resonance imaging were used as a reasonable substitution because of the vocal tract similarity for different subjects [23], [24]. In this work, all five Chinese fricatives were selected, i.e., Pinyin /f/, /s/, /sh/, /x/, /h/, as [f], [s], [ʂ], [ɕ], [x] in International Phonetic Alphabet. Using the same method in Wu's work [19], the truncated vocal tract length and EL source location were estimated at round 5 cm and 8 cm from the glottis in the imposed vocal tract shapes. Accordingly, the VTTF of the intact vocal tract and the truncated vocal tract were computed based on the 1-D digital waveguide model [25]. Fig. 2(a) shows typical VTTFs for the fricative /s/ respectively based on an integrated vocal tract and a truncated vocal tract. Two energy deficiency regions centered at 2.5 kHz and 6.5 kHz can be observed in the VTTF of the laryngectomized condition, indicating two main frequency regions where the energy will be compensated. For simplicity, the entire frequency domain (0–10000 Hz) was divided into 25 successive segments (400 Hz/segment). Then, the magnitude difference between the VTTF of normal subject and laryngectomee was averaged in each frequency segment and illustrated as in Fig. 2(b). Finally, the passband edge frequencies F_L and F_H of the two main energy attenuation regions were determined. Because the energy attenuation regions were similar for different fricatives, the parameters of the compensation source were constant for all the fricatives as listed in Table II.

TABLE II
THE PARAMETERS OF THE MONOPOLE SOURCE AND THE COMPENSATION SOURCES

Source type	F_L(Hz)	F_H(Hz)	k	Filter order
Compensation source 1	2000	3500	1.5	6(BPF)
Compensation source 2	6000	8000	2.0	6(BPF)
Monopole source	1100	–	0.1	2(LPF)

The BPF and LPF represent the band-pass filter and low-pass filter. The F_L and F_H are the low cutoff frequency and the high cutoff frequency of the filters.

C. Source Combination

As shown in Fig. 1, a same Gaussian white noise was separately filtered by the spectral models of the monopole source, dipole source, and compensation source. Then, the final hybrid noise source was the weighted sum of the three types of filtered waveforms as

$$u = k_m * u_m + u_d + \sum\nolimits_i k_{ci} * u_{ci} \tag{1}$$

where u_m, u_d, and u_c are the monopole source, dipole source, and compensation source respectively. The k_m and k_c are relative amplitude coefficients of the monopole source and the compensation source normalized by the amplitude of the dipole source in the time domain.

In this study, the amplitude of the dipole source was set as a fixed constant value for all fricatives, namely, the scaling factor of the dipole source amplitude is 1.0. According to the previous study [21], the relative amplitude coefficient of the monopole source k_{m} is generally set as 0.3 in normal speech production. However, considering the low-pass effect (about 10 dB in 0–1000 Hz) of the neck tissue in EL speech production [25], the k_m was set to be 0.1 in this model.

In addition, the relative amplitude coefficient of the compensation source k_c was determined as

$$k_c = 10^{(cMD - pMD)/20} \tag{2}$$

where *cMD* is the magnitude difference that is expected to be compensated in the energy attenuation region as shown in Fig. 2(b) and 3, and *pMD* is the magnitude difference of the dipole source u_d and the compensation source u_c in the same region as shown in Fig. 3. In this study, the k_c was set as 1.5 for the first compensation source at 2.5–3.5 kHz and 2.0 for the second compensation source at 6.0–8.0 kHz as listed in Table II.

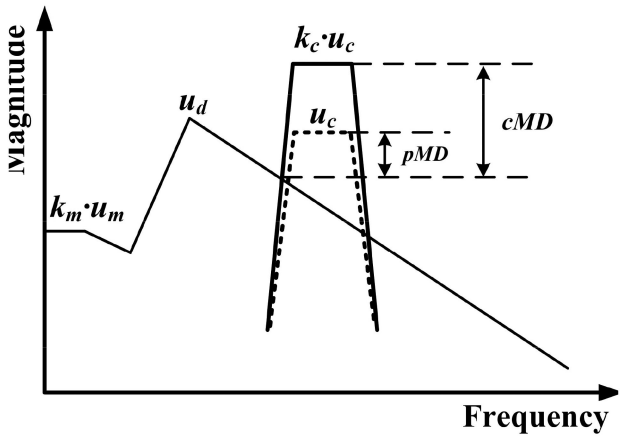

Fig. 3. The sketch of the scaling factor calculation in the frequency domain. The u_m, u_d, and u_c are the monopole source, dipole source, and compensation source. The k_m and k_c are the amplitude coefficients to modify the spectral magnitudes of the monopole source and compensation source with respect to that of the dipole source.

III. Experiments

Twenty-one male laryngectomees participated in the experiment of Mandarin fricative reconstruction. The patients were averaged 63.8 (SD = 5.7) years old. All of the laryngectomees were native Mandarin speakers and used the EL as their major communication manners. They had no history of speech problems except due to laryngectomy. Meanwhile, twenty-one healthy male subjects also participated in the normal fricatives recordings. The healthy subjects are also native Mandarin speakers and averaged 62.7 (SD = 8.7) years old, and they had no history of speech problems.

During the experiment, the subjects were seated at a comfortable posture, holding a linear vibrator (Brüel & Kjær, Model 4810, Skodsborgvej, Denmark) against the neck like using a commercial EL. Then, the laryngectomees tried to produce the Mandarin fricatives with the corresponding hybrid noise sources and the wideband noise source provided by the linear vibrator, respectively. In addition, the EL speech produced by the laryngectomees using a commercial EL (Xiwang VII) and normal speech produced by healthy subjects were also recorded. The speech material is a text entitled 'Beifeng he Taiyang'. It is 170 words long, contained all the Mandarin fricatives and vowels [26]. Therefore, all laryngectomees produced EL fricatives four times by using conventional EL source for all fricatives, wideband noise source for all fricatives, hybrid noise source for /x, s, sh/ and hybrid noise source for /h, f/ respectively. The excitation sources were exclusively used and controlled manually based on the pronunciation onset/offset while producing the speech materials. At the same time, the speech signals were collected by a microphone placed 10 cm in front of the mouth and recorded at a sampling rate of 44100 Hz with 16-bit quantization.

All the recordings were analyzed from aspects of acoustic characteristics and perceptual intelligibility to evaluate the improvement of the EL fricatives using different sources. In the acoustic analysis, time-domain overall amplitude and spectral features, such as energy ratio between high-frequency region and low-frequency region (H/L ratio), spectral moments that majorly reflect the acoustic characteristics of fricatives were investigated [27]. As in Maniwa and Jongman's work [27], the speech signals were firstly pre-processed with a high-pass filter (cutoff frequency 300 Hz). Then, the power spectra of the fricatives were estimated by AR model (order = 150) with a 40 ms hamming window located at the middle of each fricative. The H/L ratio, calculated as the ratio between the energy above 4 KHz and in lower frequency region, plays a role in the perception intelligibility of alaryngeal speech [28]. The first, third and fourth spectral moments (representing spectral mean, skewness, and kurtosis respectively) that capture both local and global spectral information were computed following the procedures described by Forrest [29].

In the perceptual test, the intelligibility of the EL fricatives was evaluated. There were three sets of listening materials used in the experiment, and each set had the syllables with fricative-vowel structure. To avoid the vowel influence, the vowel in all syllables was the same one, i.e., vowel /a/ in Pinyin as [a] in IPA. In each testing text, every fricative was repeated by 42 times (2 times/person $\times$ 21 person). In the first set, the syllables were the recorded EL speech produced using the commercial EL. Two other sets were obtained by replacing the fricatives in the first set with the EL fricatives produced using the hybrid noise sources and wideband noise source respectively, without changing the vowel and voice onset time.

Ten healthy subjects averaged 26.0 (SD = 1.5) years old participated in the listening test. All the listeners were native Chinese speakers and highly educated. They all had no history of hearing disorders. To avoid learning and experience effects, the order of the syllables was set randomly. The listeners were seated in a quiet room (environmental noise smaller than 30 dB) and presented with the speech stimuli by a loudspeaker (placed 1.0 m away from the listeners) in about 60 dB (detected 10 cm away from the loudspeaker) that is similar with the speech level of EL speech. The listeners were instructed to transcribe the fricatives using broad phonetic transcription, in which the listeners were instructed to write down the syllables they heard without any given options. The listeners were not privy to the experimental paradigm and purpose. The intelligibility score was calculated as the mean percentage of correct responses to fricatives.

IV. Results

A. Acoustical Characteristics

1) Overall Amplitude: Fig. 4 shows the overall amplitudes of the EL fricatives produced by hybrid noise source (abbreviated as ELF-HNS) and wideband noise source (abbreviated as ELF-WNS) undergone the same excitation amplitude. Above all, the overall amplitude of ELF-HNS is about 3 dB larger than the overall amplitude of ELF-WNS on average. Then, each ELF-HNS also has significantly larger overall amplitude than each ELF-WNS (one-way ANOVA, $p < 0.05$ for each fricative). This indicates that the hybrid noise source has lower energy

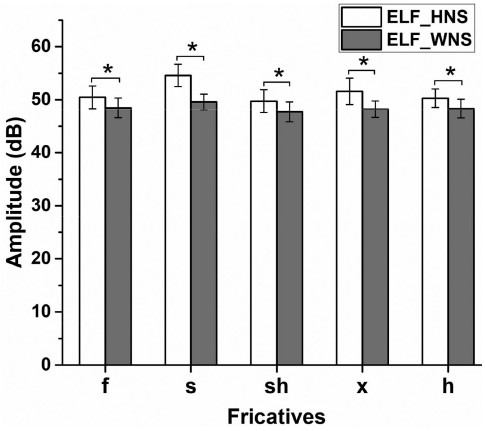

Fig. 4. The overall amplitudes of EL fricatives produced by hybrid noise source (ELF-HNS) and wideband noise source (ELF-WNS). $*$ represents significant difference (one-way ANOVA, $p < 0.05$).

attenuation across the neck and vocal tract in EL fricative production than the wideband noise source.

2) Spectral Shape: Fig. 5 shows the power spectra of the EL fricatives produced by hybrid noise source, wideband noise source and conventional EL source (abbreviated as CELF), respectively. Firstly, both the ELF-HNS and ELF-WNS have obviously different spectral shapes from the CELF. The spectra of ELF-HNS and ELF-WNS can be described as a major broad peak around 5 KHz carrying many minor sharp characteristic peaks. These are roughly consistent with the spectral shapes of English voiceless fricatives [27]. However, the CELF only remain the minor sharp characteristic peaks, missing the major broad peaks around 5 KHz. Secondly, it is visible that the ELF-HNS and ELF-WNS have smaller energy in low-frequency region (lower than 4 KHz), but larger energy in high-frequency region (higher than 4 KHz) than CELF. Quantificationally, Fig. 6 shows the ratio between the energy above 4 KHz and the energy in the lower frequencies (H/L ratio) of EL fricatives produced by different sources. All the H/L ratios of normal fricatives (NF), ELF-HNS and ELF-WNS are larger than the H/L ratios of CELF. The H/L ratio differences with NF is 4.0 dB for ELF-HNS and 4.6 dB for ELF-WNS on average, that are both much smaller than the H/L ratio difference between CELF and NF (11.2 dB on average). Thirdly, the energy defects between 6–8 KHz that are obvious in ELF-WNS are also partly compensated in the ELF-HNS, indicating the compensation source effective on eliminating the energy defects caused by abnormal VTTF in EL speech production. Qualitatively, these results indicate that both the hybrid noise source and wideband noise source can effectively reconstruct main spectral shape characteristics of EL fricatives and clearly reduce the excessive low-frequency energy of the EL fricatives, however, the hybrid noise source has an advantage over the wideband noise source in compensating the energy defects in EL fricative production.

3) Spectral Moments: Fig. 7 compares the first, the third and the fourth spectral moments (representing mean frequency, skewness and kurtosis respectively) of the EL fricatives produced by different sources. For the spectral mean, the mean frequency of ELF-HNS (4930 Hz on average) resembles the

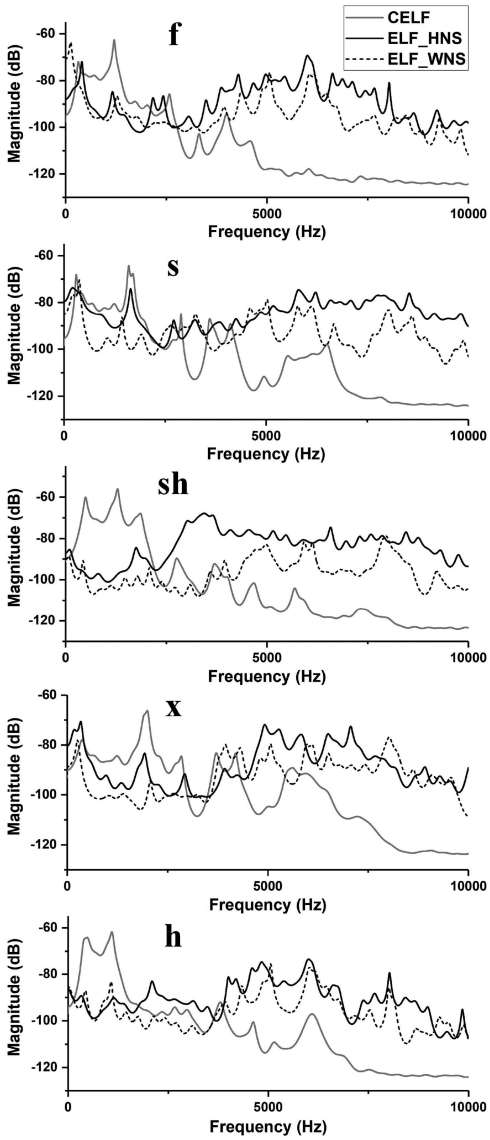

Fig. 5. Spectral shapes of EL fricatives produced by one laryngectomee with the hybrid noise source, the wideband noise source and the conventional EL source respectively. The ELF-HNS is the fricative produced by the hybrid noise source (black solid lines). The ELF-WNS is the fricative produced by wideband noise source (dash lines). CELF is the fricative produced by conventional EL source (gray solid lines).

mean frequency of normal fricatives (4980 Hz on average) that is slightly larger than that of ELF-WNS (4740 Hz on average) but much larger than that of CELF (2750 Hz on average). In addition, as shown in Fig. 7(a), for each fricative, the CELF also has significantly smaller spectral means than the ELF-HNS, ELF-WNS and NF (one-way ANOVA, $p < 0.05$ for each case). This indicates that, compared with the CELF, the average energy concentration of the ELF-HNS and ELF-WNS are at relatively higher frequencies that approach those of normal fricatives, which also can be inferred from the results shown in Fig. 5 and Fig. 6.

Fig. 7(b) shows the skewness of NF and EL fricatives produced by different sources. The skewness of ELF-HNS, CELF and NF are positive and the skewness of ELF-WNS is negative.

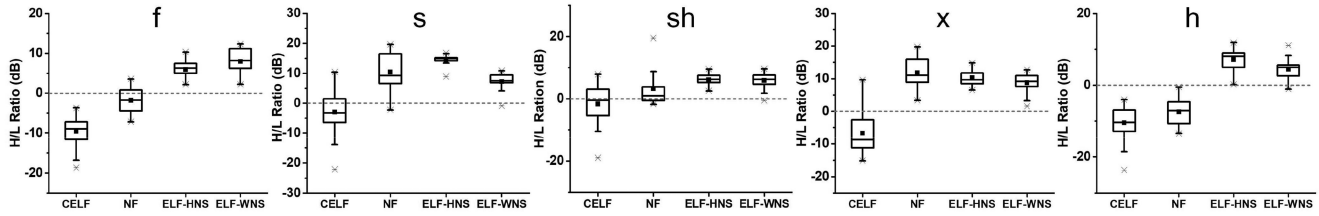

Fig. 6. The high-frequency/low-frequency energy ratio (critical frequency 4 KHz) of fricatives produced by different sources. CELF is EL fricatives produced by conventional EL source; NF is normal fricatives; ELF-HNS is EL fricatives produced by hybrid noise source and ELF-WNS is EL fricatives produced by wideband noise source.

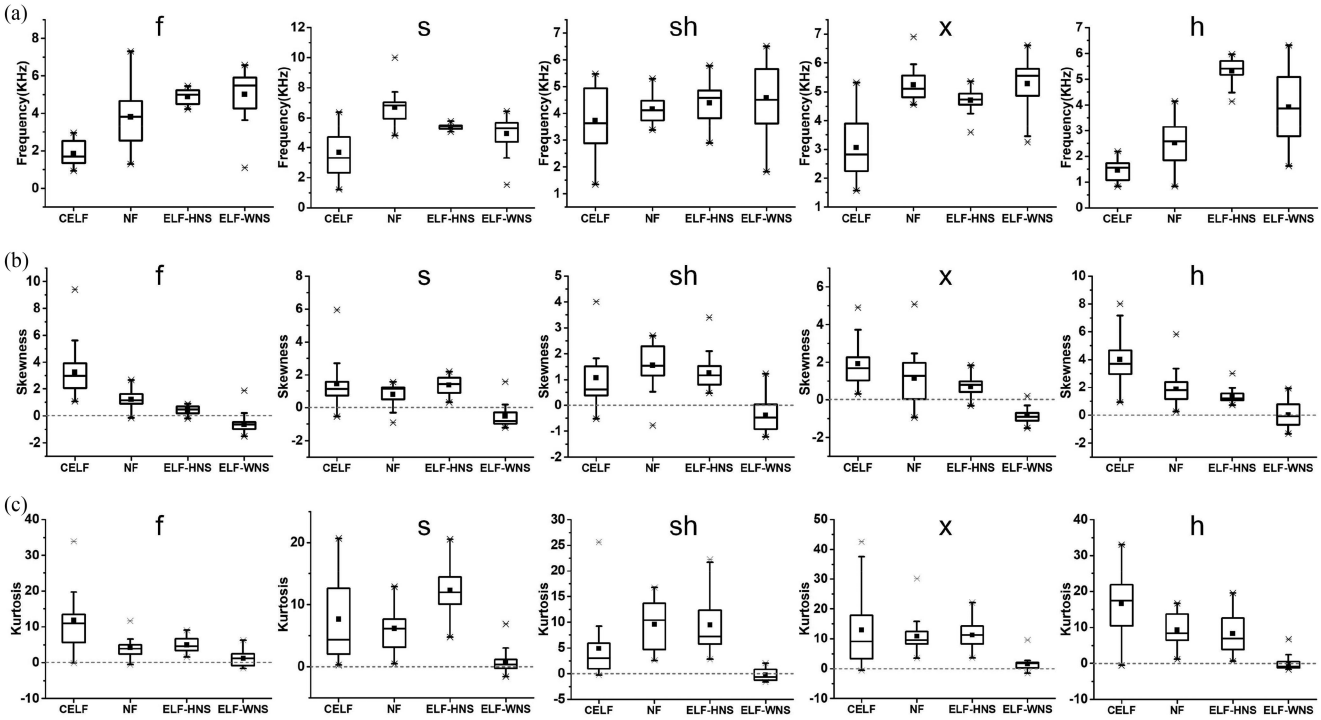

Fig. 7. Spectral moments of EL fricatives produced by different sources. (a) The first spectral moment: mean frequency; (b) the third spectral moment: skewness; (c) the fourth spectral moment: kurtosis. CELF is EL fricatives produced by conventional EL source; NF is normal fricatives; ELF-HNS is EL fricatives produced by hybrid noise source and ELF-WNS is EL fricatives produced by wideband noise source.

This result implies negative spectral tilts with energy bias in the lower frequencies from mean frequency for ELF-HNS, CELF and NF, whereas positive spectral tilts with energy bias in the higher frequencies from mean frequency for ELF-WNS. In addition, the difference between ELF-HNS and NF (averaged 0.3) is also much smaller than the difference between CELF and NF (averaged 1.1) or the difference between ELF-WNS and NF (averaged 1.8). These indicate that the hybrid noise source achieves much better performance on controlling the energy bias from spectral means for reconstructed fricatives than the conventional EL source and wideband noise source.

With respect to kurtosis, as shown in Fig. 7(c), the kurtoses of ELF-HNS (averaged 9.2), CELF (averaged 11.2) and NF (averaged 8.0) are significantly larger than the kurtoses of ELF-WNS (averaged 0.7) (one-way ANOVA, $p < 0.05$ for each case). The difference between ELF-HNS and NF (1.2) is also much smaller than the difference between CELF and NF (3.2) or the difference between ELF-WNS and NF (7.3). These results indicate that all ELF-HNS, CELF and NF have more peaked energy distributions (indicating the major energy concentrated in a narrower band) than ELF-WNS, however, the peakedness of the energy distribution of ELF-HNS is closest to those of normal fricatives.

In summary, compared with conventional EL source, both hybrid noise source and wideband noise source can obviously improve the average energy concentration of reconstructed fricatives to higher frequencies, approaching normal fricatives. Further, the skewness and kurtoses of fricatives produced by the hybrid noise source are closest to those of normal fricatives, however, are obviously larger than those of fricatives produced by the wideband noise source. Thus, the hybrid noise source performs much better than the wideband noise source and conventional EL source in spectral moment reconstruction of EL fricatives.

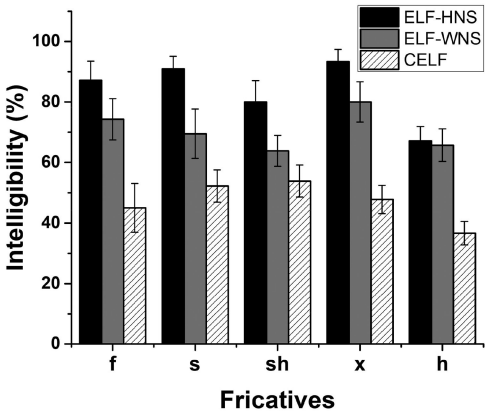

Fig. 8. Intelligibility of fricatives produced by different sources. CELF is EL fricatives produced by conventional EL source; ELF-HNS is EL fricatives produced by hybrid noise source and ELF-WNS is EL fricatives produced by wideband noise source.

B. Perceptual Intelligibility

Fig. 8 shows the perceptual intelligibility of the ELF-HNS, ELF-WNS and CELF. The intelligibility of ELF-HNS is averaged 83.71% and the intelligibility of ELF-WNS is averaged 70.67% that are both much larger than the average intelligibility of CELF 47.11%. Moreover, for each fricative, the ELF-HNS and the ELF-WNS also have significantly larger intelligibility than the CELF (one-way ANOVA, $p < 0.05$ for each fricative in both cases). In addition, each ELF-HNS also has significantly larger intelligibility than each ELF-WNS ($p < 0.05$ for each fricative, except /h/). These results indicate that both of the hybrid noise source and wideband noise source can effectively improve the intelligibility of EL fricatives, but the hybrid noise source achieves a much larger improvement of intelligibility than the wideband noise source.

V. Discussion

The purpose of this work was to improve the intelligibility of EL fricatives through eliminating the abnormal acoustic characteristics of EL fricatives. Due to the important role of EL source in EL speech production, a hybrid noise source was designed in consideration of acoustic features of the turbulent noise sources in normal consonant production and sound transmission characteristics of the vocal tract in laryngectomized subjects, to improve acoustic quality and intelligibility of the EL fricatives.

Above all, both of the hybrid noise source and wideband noise source can effectively improve the intelligibility of EL fricatives. Previous study reported that the "turning voiceless into voiced" is the major cause of low intelligibility for EL speech [8], [30]. This is because laryngectomees can only produce extremely weak consonant signals due to the lack of airflow to produce strong noise burst and noise frication after laryngectomy [2]. However, the EL can only provide a periodic vibration signal that resembles the vowel voice source. Therefore, the conventional EL fricatives are easily perceived as voiced consonants. The hybrid noise source and wideband noise source are both noise signals, resembling the normal fricative sources that are turbulent noise. Expectedly, adding a noise source will improve the intelligibility of EL voiceless consonants. In addition, the average energy concentration of ELF-HNS and ELF-WNS is also closer to the average energy concentration of normal fricatives than that of CELF (See Fig. 7a). Resembling the vowel source, the conventional EL source is characterized by average energy concentration in low-frequency region [31]. However, the wideband noise source has flat energy distribution over the spectrum and the hybrid noise source have average energy concentration around 5 KHz, thus the ELF-HNS and ELF-WNS have larger energy concentration frequencies than CELF (See Fig. 7a). Miller and Nicely [32] reported that the acoustic characteristics of normal fricatives dominating the perceptual intelligibility are in the frequency region of higher than 3 KHz. Therefore, for the both reasons, the ELF-HNS and ELF-WNS have larger intelligibility than CELF.

Furthermore, the hybrid noise source also achieves higher intelligibility for reconstructed fricatives than the wideband noise source. This is mainly attributed to better spectral characteristics achieved by the hybrid noise source. Firstly, the hybrid noise source is designed as a broad peak on spectrum that approaches the spectral characteristics of normal fricative source [18]. However, the wideband noise source is flat on spectrum. Therefore, the ELF-HNS has more peaked energy distribution (indicating higher kurtosis, see Fig. 7c) than ELF-WNS, approaching the normal fricatives. Secondly, the spectral tilt of the hybrid noise source are also designed based on the spectral tilt of normal fricative sources. Thus, both the ELF-HNS and NF are characterized by energy bias in lower frequencies from mean frequencies (indicating positive skewness, see Fig. 7b). Ignoring the spectral tilts, the wideband noise source leads to energy bias in higher frequencies from mean frequencies (indicating negative skewness, see Fig. 7b), which is contrary to the energy bias of ELF-HNS and NF. Thirdly, the hybrid noise source also effectively compensates the energy deficiency region in EL fricatives, but the wideband noise source is unable to achieve this compensation (See Fig. 5). This indicates that the compensation source in hybrid noise source indeed works well at eliminating the energy defects caused by truncated vocal tract. From above, achieving higher skewness and kurtosis to approach normal fricatives, and compensating the energy deficiency in EL fricatives, the hybrid noise source get significantly larger intelligibility improvement for reconstructed EL fricatives than wideband noise source.

Except for achieving higher intelligibility, the hybrid noise source also have smaller energy attenuation than wideband noise source in EL fricative production (see Fig. 4). This is because the vocal tract transfer functions of most fricatives can be approximately described as high-pass filters with cutoff frequency in high-frequency region [33], [34]. Therefore, the energy concentration in high-frequency region of hybrid noise source benefits the energy transmission through the vocal tract.

In summary, this work succeeded in improving the intelligibility of the EL fricatives by ameliorating the abnormal spectral features of the EL source. Although both of the hybrid noise source and wideband noise source can effectively reconstruct intelligible Mandarin fricatives for laryngectomy, the hybrid noise source has larger energy transmission and can achieve

significantly larger intelligibility improvement than wideband noise source by achieving better spectral characteristics. Thus, the hybrid noise source is a more feasible and satisfied approach for reconstructing EL fricatives than the wideband noise source.

At present, this study successfully designs an appropriate source for reconstructing EL fricatives. The hybrid noise source is a key point and the foundation for developing an EL system that can produce both consonants and vowels. In previous work, we had developed an experimental EL system that can realize automatic on/off control and output specific voice sources for corresponding vowel phonemes via lip deformation [12], [19], [35]. Based on the experimental EL system, the future projects will focus on the controlling of driving the consonant sources by physiological signals (such as EMG) and combination of the vowel and consonant sources to produce an integrated syllable. In addition, for simplifying the source model, many parameters of synthesizing hybrid noise sources are set fixed, ignoring some special characteristics of different fricatives. Probably, flexible parameter settings can achieve better results, but will also lead to increase of the model complexity. Even so, the simplified hybrid noise source still can achieve high intelligibility for reconstructed fricatives. On the other hand, the hybrid noise is also a good choice for alaryngeal speech post-processing to improve the EL speech intelligibility, such as alaryngeal speech repair and alaryngeal speech coding/decoding based on analysis-synthesis process.

VI. Conclusion

To improve the poor intelligibility of Mandarin EL fricatives, this study proposed a hybrid noise source that considers the spectral features of normal fricative sources and eliminates the abnormal spectral features caused by the truncated vocal tracts and improper voice source locations. Compared with the wideband noise source and the conventional EL source, the hybrid noise source obviously improves the spectral properties of reconstructed EL fricatives. Firstly, the global spectral shapes of EL fricatives produced by hybrid noise source (ELF-HNS) and wideband noise source (ELF-WNS) are basically in consistent with those of normal fricatives (NF), however, are much better than those of conventional EL fricatives (CELF). Secondly, the average energy concentrations of ELF-HNS and ELF-WNS are similar with those of NF, but are in significantly higher frequencies than those of CELF. Thirdly, the spectral skewness and spectral kurtoses of ELF-HNS are mostly in line with those of normal fricatives, however, are much larger than those of ELF-WNS. In addition, the hybrid noise source also have larger transmission efficiency in EL speech production than the wideband noise source. Achieving basic spectral shapes of fricatives, both the hybrid noise source and wideband noise source largely improve the intelligibility of EL fricatives. However, benefiting from better spectral characteristics, the hybrid noise source achieves a significantly larger improvement of intelligibility for EL fricatives than the wideband noise source. These results indicate that the hybrid noise source is a feasible and effective method of substituting the conventional EL source in the EL fricative production, which is much better than the wideband noise source. Besides, the hybrid noise source is potential to further improve the intelligibility of EL consonants by extending to other Mandarin consonants, such as affricates.

References

[1] W. Chen *et al.*, "Cancer statistics in China, 2015," *CA Cancer J. Clin.*, vol. 66, no. 2, pp. 115–132, 2016.

[2] K. E. V. Sluis, L. V. D. Molen, R. J. J. H. V. Son, F. J. M. Hilgers, P. A. Bhairosing, and M. W. M. V. D. Brekel, "Objective and subjective voice outcomes after total laryngectomy: a systematic review," *Eur. Arch. Oto-Rhino-Laryngol.*, vol. 275, no. 1, pp. 11–26, 2018.

[3] S. Xi, "Effectiveness of voice rehabilitation on vocalisation in postlaryngectomy patients: a systematic review," *Int. J. Evidence-Based Healthcare*, vol. 8, no. 4, pp. 256–258, 2010.

[4] A. M. Pou, "Tracheoesophageal voice restoration with total laryngectomy," *Otolaryngologic Clin. North Amer.*, vol. 37, no. 3, pp. 531–545, 2004.

[5] R. E. Hillman, M. J. Walsh, G. T. Wolf, S. G. Fisher, and W. K. Hong, "Functional outcomes following treatment for advanced laryngeal cancer. Part I–Voice preservation in advanced laryngeal cancer. Part II–Laryngectomy rehabilitation: The state of the art in the VA System. Research Speech-Language Pathologists. Department of Veterans Affairs Laryngeal Cancer Study Group," *Ann. Otol. Rhinol. Laryngol. Supplement*, vol. 172, pp. 1–27, 1998.

[6] H. Liu and M. L. Ng, "Electrolarynx in voice rehabilitation," *Auris Nasus Larynx*, vol. 34, no. 3, pp. 327–332, 2007.

[7] R. Kaye, C. G. Tang, and C. F. Sinclair, "The electrolarynx: Voice restoration after total laryngectomy," *Med. Dev.*, vol. 10, pp. 133–140, 2017.

[8] G. S. Meltzner and R. E. Hillman, "Impact of aberrant acoustic properties on the perception of sound quality in electrolarynx speech," *J. Speech Lang. Hear. Res.*, vol. 48, no. 4, pp. 766–779, 2005.

[9] K. F. Nagle, T. L. Eadie, D. R. Wright, and Y. A. Sumida, "Effect of fundamental frequency on judgments of electrolaryngeal speech," *Amer. J. Speech Lang. Pathol*, vol. 21, no. 2, pp. 154–166, 2012.

[10] Y. Y. Qi and B. Weinberg, "Low-frequency energy deficit in electrolaryngeal speech," *J. Speech Hear. Res.*, vol. 34, no. 6, pp. 1250–1256, 1991.

[11] Y. Saikachi, K. N. Stevens, and R. E. Hillman, "Development and perceptual evaluation of amplitude-based F0 control in electrolarynx speech," *J. Speech Lang. Hear. Res.*, vol. 52, no. 5, pp. 1360–1369, 2009.

[12] C. Wan, E. Wang, L. Wu, S. Wang, and M. Wan, "Design and evaluation of an electrolarynx with tonal control function for Mandarin," *Folia Phoniatrica Et Logopaedica*, vol. 64, no. 6, pp. 290–296, 2012.

[13] L. Wang, Z. Qian, Y. Feng, and H. Niu, "Design and preliminary evaluation of electrolarynx with F0 control based on capacitive touch technology," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 3, pp. 629–636, Mar. 2018.

[14] S. K. Basha and P. C. Pandey, "Real-time enhancement of electrolaryngeal speech by spectral subtraction," in *Proc. Commun.*, 2012, pp. 1–5.

[15] H. J. Niu, M. X. Wan, S. P. Wang, and H. J. Liu, "Enhancement of electrolarynx speech using adaptive noise cancelling based on independent component analysis," *Med. Biol. Eng. Comput.*, vol. 41, no. 6, pp. 670–678, 2003.

[16] K. Xiao, "The defect in the rebuilding of electrolarynx speech," BA theses, Dept. Biomed. Eng., School Life Sci. Technol., Xi'an Jiaotong Univ., Xi'an, China, 2008.

[17] L. Wu, K. Xiao, W. Supin, and W. Mingxi, "Acoustic influence of the neck tissue on Mandarin voiceless consonant production of electrolaryngeal speech," *Speech Commun.*, vol. 87, pp. 31–39, 2017.

[18] S. Narayanan and A. Alwan, "Noise source models for fricative consonants," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 328–344, May 2000.

[19] L. Wu, C. Wan, S. Wang, and M. Wan, "Improvement of electrolaryngeal speech quality using a supraglottal voice source with compensation of vocal tract characteristics," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 7, pp. 1965–1974, Jul. 2013.

[20] L. R. Rabiner and R. W. Schafer, *Theroy and Applications of Digital Speech Processing*. Upper Saddle River, NJ, USA: Pearson, 2011.

[21] P. Birkholz and D. Jackél, "Noise sources and area functions for the synthesis of fricative consonants," *Rostocker Informatik Berichte*, vol. 30, pp. 17–30, 2006.

[22] S. Narayanan and A. Alwan, "Parametric hybrid source models for voiced and voiceless fricative consonants," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1996, pp. 377–380, vol. 1.
[23] B. H. Story, "A parametric model of the vocal tract area function for vowel and consonant simulation," *J. Acoust. Soc. Amer.*, vol. 117, no. 5, pp. 3231–3254, 2005.
[24] D. Beautemps, P. Badin, and R. Laboissière, "Deriving vocal-tract area functions from midsagittal profiles and formant frequencies: A new model for vowels and fricative consonants based on experimental data," *Speech Commun.*, vol. 16, no. 1, pp. 27–47, 1995.
[25] M. Karjalainen, "1-D digital waveguide modeling for improved sound synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2002, vol. 2, pp. 1869–1872.
[26] H. Liu, M. Wan, and S. Wang, "Features of listeners affecting the perceptions of Mandarin electrolaryngeal speech," *Folia Phoniatrica Et Logopaedica Official Organ Int. Assoc. Logopedics Phoniatrics*, vol. 57, no. 1, pp. 9–19, 2005.
[27] K. Maniwa, A. Jongman, and T. Wade, "Acoustic characteristics of clearly spoken English fricatives," *J. Acoust. Soc. Amer.*, vol. 125, no. 6, pp. 3962–3973, 2009.
[28] H. J. Shim, H. R. Jang, H. B. Shin, and D. H. Ko, "Spectral and Cepstral analyses of Esophageal speakers," *Archiv Für Hygiene Und Bakteriologie*, vol. 6, no. 2, pp. 47–54, 2014.
[29] K. Forrest, G. Weismer, P. Milenkovic, and R. N. Dougall, "Statistical analysis of word-initial voiceless obstruents: Preliminary data," *J. Acoust. Soc. Amer.*, vol. 84, no. 1, pp. 115–123, 1988.
[30] M. S. Weiss and A. G. Basili, "Electrolaryngeal speech produced by laryngectomized subjects: perceptual characteristics," *J. Speech Hear. Res.*, vol. 28, no. 2, pp. 294–300, 1985.
[31] F. Chen, L. L. Wong, and E. Y. Wong, "Assessing the perceptual contributions of vowels and consonants to Mandarin sentence intelligibility," *J. Acoust. Soc. Amer.*, vol. 134, no. 2, pp. EL178–EL184, 2013.
[32] G. A. Miller and P. A. Nicely, "An analysis of perceptual confusions among some English consonants," *J. Acoust. Soc. Amer.*, vol. 27, no. 2, pp. 338–352, 1955.
[33] P. Badin, "Acoustics of voiceless fricatives: Production theory and data," *STL-QPSR*, vol. 3, pp. 33–55, 1989.
[34] C. H. Shadle, "*The acoustics of fricative consonants*," Ph.D. dissertation, MIT, Cambridge, MA, USA, 1985.
[35] C. Wan, L. Wu, H. Wu, S. Wang, and M. Wan, "Assessment of a method for the automatic on/off control of an electrolarynx via Lip deformation," *J. Voice*, vol. 26, no. 5, pp. 674.e21–674.e30, 2012.

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