Radiated Noise Suppression for Electrolarynx Speech Based on Multiband Time-Domain Amplitude Modulation

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

***Abstract*—Radiated noise severely degrades the electrolarynx (EL) speech. It cannot be thoroughly suppressed by conventional frequency-domain enhancement methods. In this paper, a new method, called multiband time-domain amplitude modulation (MTAM), is proposed to reduce the radiated noise of EL speech. In the proposed method, the speech components changing slowly that represent the radiated noise are removed by directly modulating the time-domain amplitudes in multiple frequency bands. The EL speech enhanced by the proposed MTAM and the conventional frequency-domain enhancement methods (spectral subtraction and Wiener filtering) are evaluated on both acoustic and perceptual characteristics. The acoustic analysis reveals that the MTAM not only can reduce the radiated noise more thoroughly but can also easily control the residual noise intensity by adjusting a modulation parameter λ. Moreover, the MTAM can avoid causing new artificial noise that cannot be avoided by the conventional frequency-domain enhancement methods. The perceptual analysis indicates that the MTAM also have better performance on increasing the acceptability and the consonant intelligibility of EL speech than spectral subtraction and Wiener filtering. These findings validate that the MTAM indeed works well in suppressing the radiated noise of EL speech and avoiding the artificial noise.**

***Index Terms*—Electrolarynx speech, enhancement, radiated noise, speech quality, time-domain amplitude modulation.**

I. Introduction

LARYNGECTOMY is the most effective treatment for larynx cancers and is widely used all over the word, especially in China where over 250000 Chinese patients are diagnosed with larynx cancer each year [1]. The electrolarynx (EL) is a device that provides periodic mechanical vibration signal to replace the vocal fold vibration for the laryngectomees to reconstruct intelligible speech. The EL has been the most widely used method of speech rehabilitation for laryngectomees due to the advantages of easy learning, easy operating and continuous output [2]–[4]. Unfortunately, the EL speech quality and intelligibility are severely damaged by the strong noise presented in the EL speech, especially the radiated noise [5], [6]. In addition, the strong noise in the EL speech also creates difficulties for EL speech processing, such as voice activity detection and EL speech recognition. Thus, noise reduction is imperative to improve the quality and the intelligibility of EL speech.

The radiated noise in EL speech is produced by a part of energy produced by EL that is unable to pass through the neck tissue and is radiated directly into the outside environment. In contrast to the natural background noise, the radiated noise is still prominent even in a quiet environment. Previous researches have revealed that the radiated noise of EL speech is about 20–25 dB when the mouth is closed and varied over 4–15 dB across subjects for using a same EL device [7], [8]. The masking effects of radiated noise considerably contribute to the unnaturalness and the poor intelligibility of EL speech, especially for some EL consonants that are almost totally submerged by radiated noise. Therefore, suppressing the radiated noise of EL speech is essential for enhancing the EL speech.

Several researchers have attempted to suppress the radiated noise of EL speech. Norton and Bernstein [9] surrounded the EL with a one-inch-thick foam shield to reduce the radiated noise, resulting in no obvious improvement of EL speech intelligibility and quality. Many researchers have diverted their efforts to the signal processing techniques for EL speech enhancement. The spectral subtraction (SS) is the most common method for enhancing the EL speech, especially for single channel speech enhancement [8], [10]–[13]. The SS method subtracts the noise components directly in frequency domain. In fact, it is difficult to precisely estimate the noise spectrum due to its instability, which will lead to random interval errors after spectral subtraction, causing residual artificial noise by inverse Fourier transform. This residual artificial noise, also called "musical noise", is sometimes more annoying than the original noise. Although several studies have provided solutions to reduce the musical noise [10], [14]–[16], results obtained by these methods have suggested that the improvement is still limited, especially under low signal-to-noise ratio (SNR).

In order to avoid the artificial noise caused by the errors between the estimated noise and the real noise spectra after inverse Fourier transform, this paper attempts to estimate and reduce the radiated noise directly in time domain. It has been revealed that the temporal changes convey most of the linguistic information

Manuscript received January 13, 2018; revised May 2, 2018; accepted May 3, 2018. Date of publication May 9, 2018; date of current version May 25, 2018. This work was supported by the National Natural Science Foundation of China under Grants 81771854, 11404256, and 11274250. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Andy W. H. Khong. *(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.2018.2834729

of speech, dominating the perceptual intelligibility of speech, because the temporal changes reflect the movement of the vocal tract [17]. The time-domain amplitude of radiated noise changes slowly over a long duration, and can be considered as a stable signal in a limited duration in case of a patient using the same EL, because the intensity of the radiated noise is only related to the instrument itself and the users. The temporal changes of radiated noise are much smaller than that of the desired EL speech signals. Therefore, removing the speech components changing slowly is also a feasible way of suppressing the radiated noise, without removing the useful EL speech components.

In this paper, a method of multiband time-domain amplitude modulation (MTAM) is proposed to suppress the radiated noise of EL speech by removing the slowly changing speech components in different frequency bands using time-domain amplitude modulation. The remainder of the paper is structured as follows: Section II introduces the signal processing procedure of MTAM; Section III details the experiments; Section IV and V present and discuss the EL speech enhanced by the MTAM and classical enhancement methods, respectively.

II. Noise Suppression Based on Multiband Time-Domain Amplitude Modulation

As shown in Fig. 1, the radiated noise suppression based on the MTAM is conducted by modulating time-domain amplitudes of EL speech in multiple frequency bands. In terms of MTAM, the methods of time-domain amplitude modulation and multiband division are the key points of suppressing the radiated noise of EL speech.

A. Time-Domain Amplitude Modulation

As shown in Fig. 1, the time-domain amplitude modulation is conducted by EL speech multiplying the weighting coefficients that are the ratio of the modified and the original time-domain envelopes. The modified time-domain envelope is obtained by the original time-domain envelope of EL speech minus the average time-domain envelope value of radiated noise. Therefore, it is important for the MTAM to detect the radiated noise duration, which is usually conducted by voice activity detection techniques [18]. In order to reduce the method complexity, the voice activity detection techniques are not utilized in the proposed approach to ensure the voiced/unvoiced duration of EL speech. Due to the stability of the radiated noise, the average envelope value of a short unvoiced duration can be substituted as the average envelope value of the entire radiated noise. In terms of EL speech, the time-domain envelope of noise duration (unvoiced duration) is smaller than the time-domain envelope of speech duration (voiced duration). Besides, previous study data has revealed that the percentage of unvoiced duration for speech signals is in the range of 35%–43% [19]. Therefore, the time-domain envelope values ranking in the last 20% in a speech processing frame are averaged and substituted as the time-domain envelope values of the entire radiated noise in this paper. On the other hand, a controlling parameter λ ranging from 0 to 1 is proposed to multiply with the average envelope value of the radiated noise to control the intensity of the residual

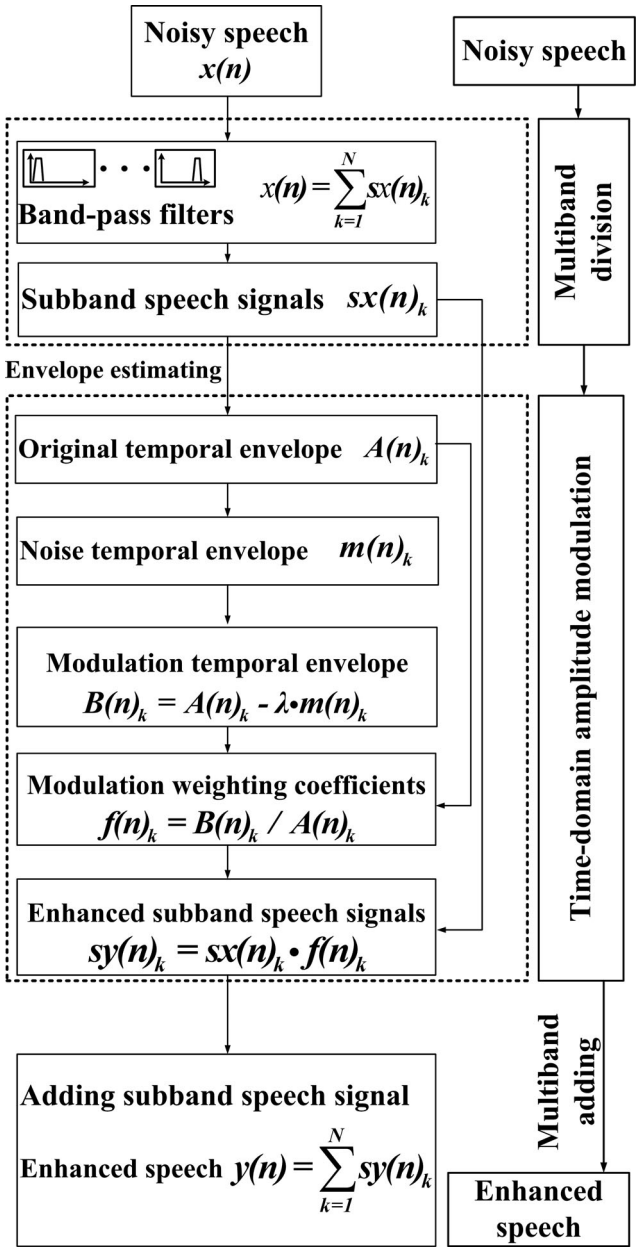

Fig. 1. Block diagram of radiated noise suppression based on MTAM method.

radiated noise. Obviously, the intensity of the radiated noise can be proportionally reduced. Larger λ will lead to smaller residual noise intensity.

B. Multiband Division

In practical application, time-domain amplitude modulation cannot be directly utilized to eliminate the radiated noise of EL speech because this operation will also subtract some energy of the desired EL speech signals at the same time, especially the weak consonants that are submerged by the strong radiated noise. The reason is that the time-domain envelope of wideband speech signal cannot reflect the time-domain envelope fluctuation of different speech components. In order to address this limitation, the time-domain envelopes extracted from different

frequency bands of EL speech are processed for more precise modulation (See Fig. 1), since the major energy of the radiated noise and the desired EL speech signals are concentrated in different frequency bands. The energy of radiated noise of EL speech majorly distributes in the range of 400-800 Hz, and the energy of EL speech mainly distributes in the range of 2-4 KHz [4]. In addition, Drullman [20], [21] revealed that the time-domain amplitude variations in successive frequency bands (bandwidth smaller than 1-oct/band) below 4 Hz can be reduced without reducing speech intelligibility for normal-hearing listeners. Therefore, the energy concentration bands of radiated noise and desired speech signal can be effectively separated by at least 5 subbands (100–400 Hz, 400–800 Hz, 800–2000 Hz, 2000–4000 Hz, larger than 4000 Hz). Undoubtedly, larger subband number leads to better separation, but also larger computational complexity. In this paper, the original EL speech is filtered into 6 bands (1-oct/band) that cover the frequency range from 100 to 6400 Hz.

C. Procedure

The processing details of MTAM are as follows:

Firstly, the noisy EL speech is divided into 6 successive frequency bands (1-oct/band) by passing through a band-pass filter bank, covering the frequency range of 100–6400 Hz:

$$x(n) = \sum_{k=1}^{6} sx(n)_k \tag{1}$$

Next, the time-domain envelopes $(A(n)_k)$ of these sub-band speech signals $(sx(n)_k)$ are estimated by Hilbert transform. Previous research has demonstrated that the low-frequency information (4–16 Hz) of the time-domain envelopes corresponding to the vocal tract movement is mainly responsible for the speech intelligibility [20], [22]. Therefore, in order to highlight the main amplitude changing information, these time-domain envelopes are smoothed by a low-pass filter (cut-off frequency is 20 Hz). Based on the original time-domain envelopes, the average time-domain envelopes of radiated noise duration (m_k) are also estimated.

Then, the modified time-domain envelopes are obtained by subtracting the average amplitude of noise duration from the original time-domain envelopes:

$$B(n)_k = \begin{cases} A(n)_k - \lambda \bullet m(n)_k, & \text{if } A(n)_k > \lambda \bullet m(n)_k \\ 0, & \text{if } A(n)_k \le \lambda \bullet m(n)_k \end{cases} \tag{2}$$

where A_k and B_k are the original and the modified time-domain envelopes, respectively, m_k is the average amplitude of the noise, and λ is the modulation parameter that controls the amplitude of the residual noise ($0 \le \lambda \le 1.0$).

Finally, the enhanced EL speech (y) is obtained by adding the sub-band speech signals multiplied by the ratios between the modified and the original time-domain envelopes:

$$y(n) = \sum_{k=1}^{6} sx(n)_k \bullet \frac{B(n)_k}{A(n)_k} \tag{3}$$

III. Experiment

A. EL Speech Signal Recording

Five male laryngectomee participated in the EL speech signal recording. The patients were 63.8 ± 5.7 years old and had undergone total laryngectomy and radiation therapy for more than 10 years. All the subjects were native Mandarin speakers and were skilled at using a commercial EL to communicate with others.

The speech material is the first paragraph of a text entitled 'Beifeng he Taiyang' (Boreas and Sun) [23]. The speech material contained all the Mandarin vowels and consonants, and was phonically and tonetically balanced for Mandarin Chinese. During the recording experiment, the patient was seated at a comfortable posture 10-cm away from a microphone. Then, the patient produced the speech materials using a commercial EL (Sevox, Servona). The EL speech was recorded at a sampling rate of 16000 Hz and digitized into 16 bits.

B. Speech Processing

Firstly, for comparative analysis, the original noisy EL speech is enhanced by the MTAM ($\lambda = 1.0$) and two classical speech enhancement methods: Wiener Filtering (WF) [24], and spectral subtraction (SS) [10]. Secondly, the parameter λ was adjusted ($0 \le \lambda \le 1$) to control the residual noise intensity during the EL speech processing in order to investigate the relationship between the signal-to-noise ratio (SNR) and the modulation parameter.

C. Perceptual Evaluation

In order to avoid the semantic association effects, only isolated words that were cut out from the speech materials were presented to the listeners in perceptual intelligibility test. The order of the syllables was randomly set to avoid learning and experience effects. The stimuli were presented in a quiet room and listeners were asked to transcript what they heard even if it was a nonsense syllable. In perceptual acceptability test, listeners were also requested to rate the four sets of speech materials (original EL speech and EL speech enhanced by MTAM, SS and WF) based on the criteria of mean opinion score (MOS) (worst:0, bad:1, poor:2, common:3, good:4, excellent:5) [7].

IV. Result

A. Acoustic Characteristics

Fig. 2 shows the typical radiated noise suppression procedure for noisy EL speech in a frequency sub-band (the third sub-band). Two main characteristics are presented. First, the radiated noise of the sub-band signal is almost fully reduced. Second, the consonant signals that are submerged by the strong radiated noise are clearly highlighted and effectively reserved after the noise reduction.

1) EL Speech Enhanced by Different Methods: Fig. 3 shows the waveforms and the spectrograms of the original noisy EL speech and the enhanced EL speech. Fig. 3(a) indicates that the radiated noise is prominent in the original EL speech and

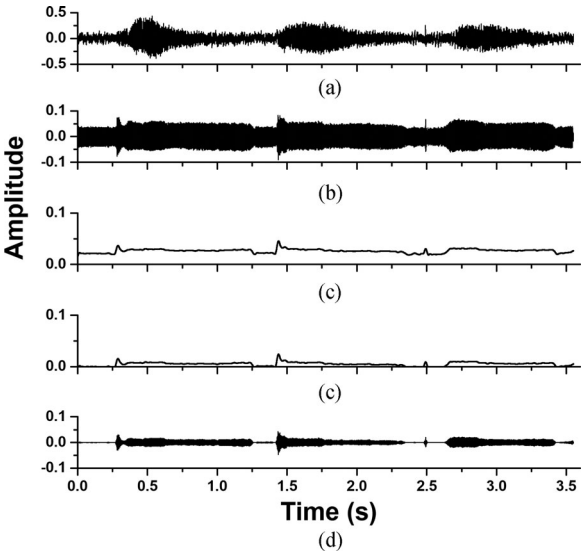

Fig. 2. Processing of EL speech in a frequency subband ($\lambda = 1.0$). (a) Original noisy EL speech; (b) EL speech of the third frequency subband; (c) Original time-domain envelope of the subband signal; (d) Modified time-domain envelope of the subband signal; (e) Enhanced EL speech of the third frequency subband.

almost submerges the weak consonant signals in time domain. It is clear that the SS, the Weiner and the MTAM methods are all able to significantly reduce the radiated noise, remaining the weak consonant signal. However, obvious residual noise can be seen in EL speech enhanced by the SS and the WF methods (See Fig. 3(b) and (c)). In contrast, no evident residual radiated noise can be seen in EL speech enhanced by the MTAM method (See Fig. 3(d)). Table I shows that the MTAM method improves the SNR of EL speech to 19.3 dB, which is much larger than the SNR improvements achieved by the SS method (12.3 dB) and the WF method (16.7 dB). Therefore, the MTAM can suppress the radiated noise more thoroughly than SS and WF.

Fig. 4 shows the power spectra of EL speech enhanced by the three methods. The smaller power spectra of enhanced EL speech indicate that all the three enhancement methods effectively reduce the energy of EL speech. The reduced energy is majorly distributed in the frequency regions without formants, especially in the low frequency region (lower than 800 Hz) that contains the main energy of the radiated noise. As shown in Fig. 4, all the three methods preserve the total energy for the first formant (F1) and the second formant (F2) and cause only a slight energy reduction for the third formant (F3) and the fourth formant (F4). In addition, it is found that the formants of other vowels are also majorly preserved by the three enhancement methods. These results indicate that all the three methods can effectively reduce the energy of radiated noise, keeping the major acoustic information preserved for EL speech. On the other hand, expect formants, the power spectrum of EL speech enhanced by the MTAM is smaller than that enhanced by SS and WF, which indicates that the MTAM can reduce more energy of radiated noise than SS and WF, leading to less residual noise.

2) Residual Noise Controlling: The MTAM method not only can reduce radiated noise thoroughly, but also can easily control the intensity of the residual noise. The SNR of EL speech enhanced by the MTAM can be calculated as followed:

$$SNR_{enhanced} = 10 \bullet \log_{10} \frac{P_{speech} - (1-\lambda)^2 \bullet P_{noise}}{(1-\lambda)^2 \bullet P_{noise}}, \quad 0 \le \lambda \le 1$$

where P_{speech} and P_{noise} are the average powers of the original EL speech and the original radiated noise, respectively.

Fig. 5 shows the SNR of EL speech enhanced by the MTAM varying as a function of the parameter λ. The increment of SNR increases as the modulation parameter λ changes from 0 to 1.0. It is found that sometimes the radiated noise cannot be thoroughly removed even by setting the λ to be 1 (See Fig. 5(a)) due to the error between the estimated and the real noise envelopes. Therefore, the parameter λ can be set larger than 1.0 to obtain more noise reduction. However, overlarge λ (larger than 1.0) will also remove some information of desired EL speech, resulting in speech distortion. In this study, it is found that the radiated noise can be well removed with the modulation parameter λ in the range of 1.0–1.2 without causing obvious speech distortion. Therefore, the modulation parameter λ in the range of 1.0–1.2 is suggested in practical applications to obtain good noise reduction. Thus, in this study, the parameter λ is always set to be 1.0.

B. Perceptual Characteristics

1) Acceptability: Fig. 6 shows the acceptability of noisy EL speech and enhanced EL speech. The original noisy EL speech has the lowest mean acceptability of 1.38, and the methods of SS, WF and MTAM improve the mean acceptability of the enhanced EL speech to 1.82, 2.21 and 2.9, respectively. There is no significant improvement of acceptability for EL speech enhanced by SS ($p > 0.05$). However, significant improvements are observed for EL speech enhanced by the MTAM method ($p < 0.001$) and the WF method ($p < 0.001$), with MTAM achieving larger improvement of acceptability than SS ($p < 0.001$) and WF ($p < 0.001$). Therefore, the MTAM has much better performance on the improvement of acceptability for enhanced EL speech than SS and WF.

2) Intelligibility of Vowels: Fig. 7 shows the mean intelligibility of vowels in EL speech enhanced by the three enhancement methods. It can be seen from the figure that all the vowels of EL speech enhanced by the three methods have considerable intelligibility (about 80%). However, there are no significant differences among the vowel intelligibility of original EL speech and the EL speech enhanced by the three enhancement methods ($p > 0.05$). Hence, reducing the radiated noise cannot significantly improve the vowel intelligibility of EL speech.

3) Intelligibility of Consonants: Fig. 8 shows the mean consonant intelligibility of EL speech enhanced by the three methods. It can be seen clearly that the consonant intelligibility of original noisy EL speech is only 52.7%, while the methods of SS, WF and MTAM have improved the mean consonant intelligibility of EL speech to 58.6%, 62.8% and 63.7%, respectively. The methods of MTAM and WF effectively improve the consonant

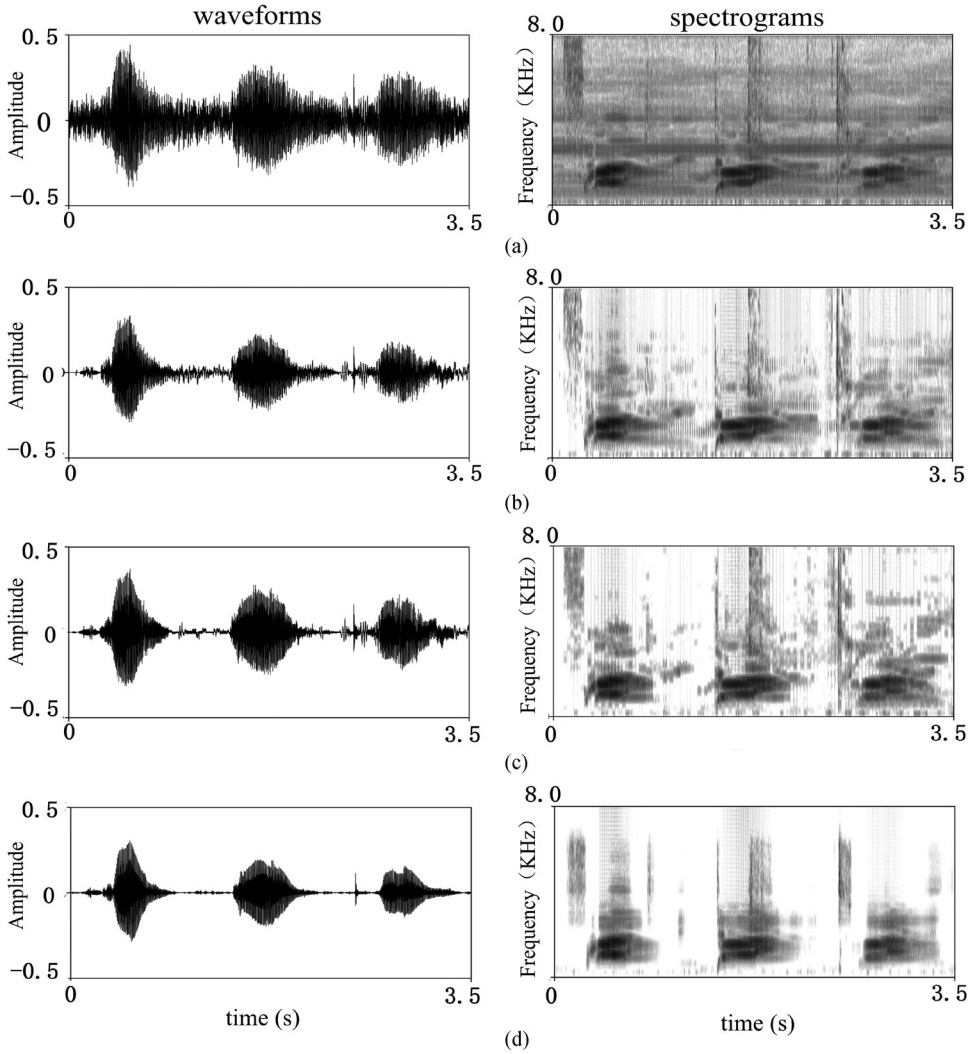

Fig. 3. Waveforms and spectrograms of EL speech before and after enhancement. The speech material is three monosyllables 'fa da ta' spoken by a laryngectomee with a commercial EL. (a) Original noisy EL speech; (b) EL speech enhanced by SS method; (c) EL speech enhanced by WF method; (d) EL speech enhanced by MTAM method ($\lambda = 1.0$).

TABLE I
SNRs OF EL SPEECH ENHANCED BY SS, WF AND MTAM

	Original SNR	SS	Weiner	MTAM
SNR(dB)	4.2	12.3	16.7	19.3

intelligibility of EL speech by 10.1% ($p < 0.001$) and 11.0% ($p < 0.001$) respectively, but there is no significant difference ($p > 0.05$) between the performance of both methods. The SS method only slightly improves the consonant intelligibility of EL speech, without significant improvement ($p > 0.05$). Thus, reducing the radiated noise by MTAM and WF can significantly improve the consonant intelligibility of EL speech, but SS fails to achieve this improvement.

Furthermore, Fig. 9 shows the perceptual accuracy of different consonant types. For original EL speech at low SNR, the aspirated plosives and fricatives have the lowest perceptual

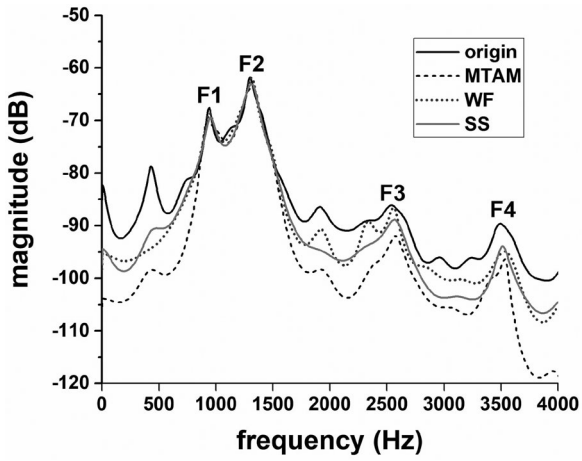

Fig. 4. Power spectra extracted from the center of /a/ produced by a laryngectomee. Black solid line represents original noisy EL speech; Black dash line represents EL speech enhanced by MTAM; Gray dot line represents EL speech enhanced by WF and gray solid line represents EL speech enhanced by SS.

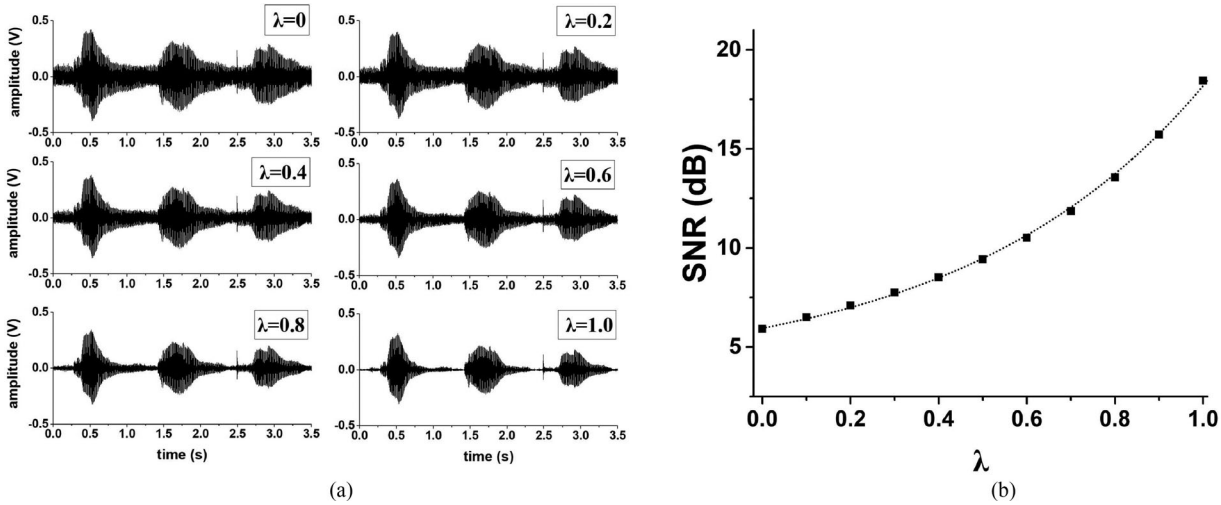

Fig. 5. SNR of enhanced EL speech as a function of the parameter λ. (a) Waveforms of enhanced EL speech with different modulation parameter; (b) SNR varying with the modulation parameter.

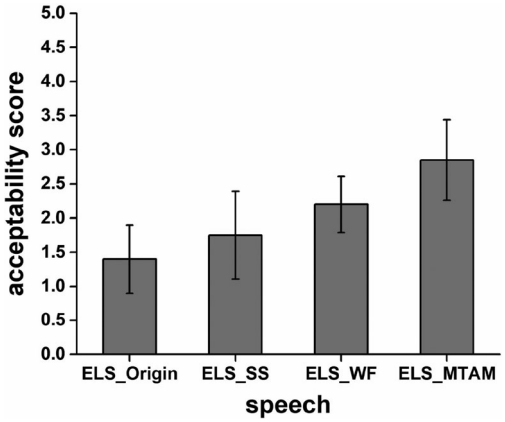

Fig. 6. Mean acceptability of EL speech before and after noise reduction. ELS_Origin represents the original noisy EL speech; ELS_SS represents EL speech enhanced by SS method; ELS_WF represents EL speech enhanced by WF method and ELS_MTAM represents EL speech enhanced by MTAM method.

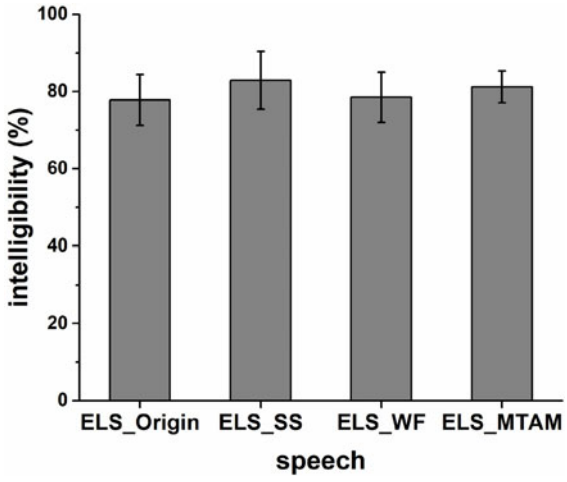

Fig. 7. Vowel intelligibility of EL speech enhanced by different methods. ELS_Origin represents the original noisy EL speech; ELS_SS represents EL speech enhanced by SS method; ELS_WF represents EL speech enhanced by WF method and ELS_MTAM represents EL speech enhanced by MTAM method.

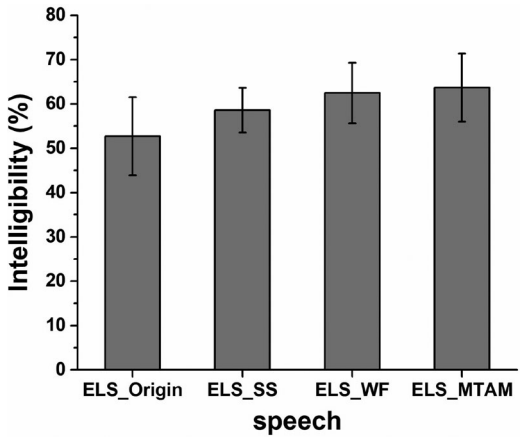

Fig. 8. Consonant intelligibility of EL speech enhanced by different methods. ELS_Origin represents the original noisy EL speech; ELS_SS represents EL speech enhanced by SS method; ELS_WF represents EL speech enhanced by WF method and ELS_MTAM represents EL speech enhanced by MTAM method.

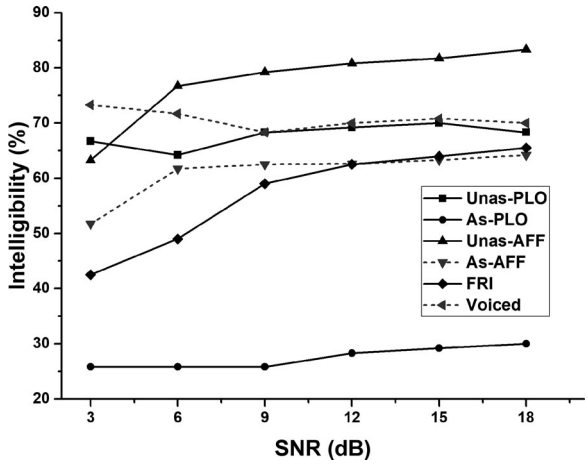

Fig. 9. Intelligibility of enhanced EL consonants as a function of pronunciation manner. Unas-PLO is the unaspirated plosive; As-PLO is the aspirated plosive; Unas-AFF is the unaspirated affricative; As-AFF is the aspirated affricative; FRI is the fricative and the voiced is the voiced consonant.

intelligibility (only 25.8% and 42.5%, respectively), contributing most to the poor intelligibility of EL speech. As the SNR of the EL speech increases to 18 dB, significant improvement of consonant intelligibility can be observed in fricatives (23%), unaspirated affricatives (20%) and aspirated affricatives (13%). The aspirated and the unaspirated plosives have small change in the intelligibility, while the intelligibility of voiced consonants is even slightly reduced as the SNR increases. These results indicate that the consonant intelligibility improvement achieved by radiated noise reduction is mainly attributed to the improvement of fricatives and affricatives.

V. Discussion

In order to improve the speech quality of EL speech, the method, called multiband time-domain amplitude modulation (MTAM), is proposed to supresses the radiated noise of EL speech. The results indicate that the MTAM method effectively reduces the radiated noise of EL speech. The EL speech enhanced by the MTAM has better performance on both acoustic and perceptual characteristics than the EL speech enhanced by spectral subtraction (SS) and Wiener filtering (WF).

Considering acoustic characteristics, the MTAM method has two advantages over SS and WF. Firstly, the MTAM method can reduce radiated noise more effectively than SS and WF without causing artificial noise. For SS and WF, the small errors between the estimated and the real noise spectra will be bound to obvious residual artificial noise such as musical noise after inverse Fourier transform. Usually, SS and WF are conducted repeatedly to obtain higher SNR. However, the artificial noise always perplexes the EL speech enhanced by SS and WF method [8]. The more unstable the noise spectrum is, the more prominent the artificial noise will be. Unlike SS and WF, MTAM method reduces the radiated noise directly in time domain by removing the signal components that change slowly. Avoiding the estimation of the noise spectrum and the inverse Fourier transform, the MTAM method can almost completely reduce the radiated noise without causing new artificial noise, even if the noise is not stable in frequency domain. Secondly, the MTAM method can easily control the intensity of residual noise for enhanced EL speech. For conventional speech enhancement method, it is unavoidable that more noise reduction leads to more speech distortion. Therefore, it is extremely important to balance the noise reduction and speech distortion. Both SS and WF are unable to precisely control the intensity of residual noise due to the difficulty of obtaining precise radiated noise estimation. However, the MTAM can directly change the time-domain amplitude of radiated noise by adjusting a modulation parameter λ, resulting in proportional reduction of radiated noise. Therefore, it is easy for MTAM to precisely control the radiated noise intensity.

The MTAM method also has better performance on perceptual characteristics of enhanced EL speech. In terms of acceptability, the EL speech enhanced by the MTAM method has higher acceptability than the EL speech enhanced by the WF and the SS methods. The reason is that the acceptability of enhanced EL speech is closely related with the residual noise. The SS method and the WF method will cause obvious artificial noise. The artificial noise caused by SS method described as warbling with tonal quality is also called musical noise that affects the acceptability of enhanced EL speech. Sometimes, this musical noise is more annoying than the original noise [24]. The residual noise caused by WF is more close to white noise that annoys the listeners less than the musical noise [25]. The MTAM method can completely reduce the radiated noise of EL speech, leading to the highest improvement of acceptability for EL speech (See Fig. 2 and Fig. 6).

In terms of intelligibility, the reduction of radiated noise cannot improve the vowel intelligibility of EL speech for all the three methods. The reason is that the energy of the vowels is much larger than the energy of radiated noise, so the radiated noise cannot mask the acoustic properties of the vowels. The vowel intelligibility of EL speech is essentially high even distorted by the radiated noise. Therefore, the speech enhancement methods cannot affect the intelligibility of vowels. In contrast, the enhancement methods can improve the consonant intelligibility of EL speech, especially the MTAM and the WF (See Fig. 8). The reason for this is that the consonant intensity is essentially much weaker than the vowel intensity and can easily be masked by the radiated noise, especially some consonants of EL speech that are produced by insufficient airflows due to the removal of throats. Thus, reducing the radiated noise can recover the consonants masked by strong radiated noise, improving the consonant intelligibility. In addition, the musical noise caused by SS method is also an annoying voice, leading to degradation for both acceptability and intelligibility. Therefore, both the MTAM method and the WF method achieve higher improvement of consonant intelligibility than the SS method.

Furthermore, it is also revealed that only the intelligibilities of fricatives and affricatives are improved by reducing the radiated noise in low SNR. The voiced consonants of EL speech are even slightly degraded by reducing the radiated noise. The fricatives and the affricatives of EL speech are extremely weak continuous noise duration and their acoustic characteristics are easily masked by the powerful radiated noise. Thus, reducing the radiated noise can remove the masking effect of radiated noise for fricatives and affricatives. In contrast, the voiced consonants are essentially periodic vibration signals and are partially similar with the radiated noise in intensity and acoustic characteristics because they are produced by a same voice source. Thus, reducing the radiated noise also leads to loss of some acoustic information for voiced consonants, causing speech degradation. For plosives, the noise burst is still prominent in the EL speech that cannot be easily masked by the radiated noise. Besides, the intelligibility of plosives is majorly affected by the voice onset time rather than the radiated noise intensity [26], [27]. Thus, EL plosives cannot be improved effectively by the reduction of radiated noise, especially the aspirated plosives. Therefore, the aspirated plosive is the choke point of improving the intelligibility of EL speech.

Considering MTAM in a single subband, the computational complexity is mainly contributed by four parts: (1) the

band-pass filtering (filter order: p) needs about $4 \cdot p \cdot N$ multiplications and $4 \cdot p \cdot N$ adds, where N is the data length; (2) the envelope estimation by Hilbert transform requires two FFT, that contributes $N \cdot \log_2 N$ multiplications and $2 \cdot N \cdot \log_2 N$ adds; (3) the noise amplitude estimation requires $N/5$ adds and 1 division; (4) the amplitude modulation needs N adds, N divisions and N multiplications. Totally, the MTAM requires about $k \cdot (N \cdot \log_2 N + 4 \cdot (p+1) \cdot N)$ multiplications, $k \cdot (2 \cdot N \cdot \log_2 N + 4 \cdot (p + 1/5) \cdot N)$ adds and $k \cdot (N+1)$ divisions, where k is the number of frequency subbands. The subtraction is considered as add approximatively in this study. With respect to SS and WF, the main computational complexity is majorly contributed by the fast Fourier transform (FFT). The SS requires about $1.5 \cdot N \cdot \log_2 N + (6 - 1.5 \cdot \log_2 L) \cdot N$ multiplications and $3 \cdot N \cdot \log_2 N + (1 - 3 \cdot \log_2 L) \cdot N$ adds, where N is the data length, and L is the number of framing. The WF requires about $1.5 \cdot N \cdot \log_2 N + (9 - 1.5 \cdot \log_2 L) \cdot N$ multiplications, $3 \cdot N \cdot \log_2 N + (1 - 3 \cdot \log_2 L) \cdot N$ adds and N divisions. By contrast, the MTAM has larger computational complexity than both SS and WF although the computational complexity of all the methods is at the same magnitude order of $O(N \cdot \log_2 N)$. Therefore, the MTAM cannot cause a rapid increase of time consumption with the data length increase, and yet it is not suitable for a real-time processing.

In summary, the MTAM method performs well at dealing with the radiated noise of EL speech without causing any artificial noise. Moreover, the MTAM method also has better performance on perceptual characteristics, especially on the acceptability, than classical SS and WF methods. In practical application, there are two aspects requiring attention while using the MTAM to suppress noises. Firstly, the MTAM is only good at suppressing the noise whose time-domain amplitude changes are much slower than that of the desired speech, such as the radiated noise of EL speech. Secondly, the MTAM method cannot provide support for real-time noise reduction, because a reliable and precise time-domain envelope of noise duration must be obtained through a relatively long processing frame (at least longer than two-syllable length suggested). Still, the MTAM method is a very good post-processing method for suppressing the radiated noise of EL speech. In addition, this method can also be used for EL speech pre-processing, providing support for other speech processing, such as voice activity detection and speech recognition.

VI. Conclusion

In this paper, a method multiband time-domain amplitude modulation (MTAM) is proposed to enhance noisy EL speech by suppressing the radiated noise without causing residual artificial noise. The comparison of the MTAM with classical spectral subtraction (SS) and Wiener filtering (WF) methods reveal that the MTAM can achieve better performance on both acoustic and perceptual characteristics. In terms of acoustic characteristics, firstly, the MTAM can completely reduce the radiated noise, causing no residual noise and artificial noise that cannot be avoided for SS and WF. Secondly, the MTAM can also easily control the intensity of radiated noise by adjusting a modulation parameter λ, which is also hard for SS and WF methods. With respect to perceptual characteristics, the acceptability of EL speech enhanced by MTAM is much better than those achieved by SS and WF. In addition, the consonant intelligibility of EL speech enhanced by MTAM is also better than that achieved by SS. Meanwhile, it is revealed that the radiated noise reduction only benefits the fricatives and the affricatives. The aspirated consonants that contribute most to the poor intelligibility of EL speech cannot be improved by reducing the radiated noise. This finding perhaps can provide some reference for further research about improving the intelligibility of EL speech. In sum, the proposed MTAM is a good method for suppressing the radiated noise of EL speech without causing residual artificial noise and can also be applied to suppress other noises that have slower temporal changes than desired signals.

References

[1] W. Q. Chen, R. S. Zheng, P. D. Baade, S. W. Zhang, H. M. Zeng, and F. Bray "Cancer statistics in China, 2015," *Ca-a Cancer J. Clinicians*, vol. 66, pp. 115–132, Mar./Apr. 2016.

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

[3] T. S. L. Verkerke and G. J., "Sound-producing voice prostheses: 150 years of research," *Annu. Rev. Biomed. Eng.*, vol. 16, pp. 215–245, 2014.

[4] H. Liu and M. W. L. Ng, "Electrolarynx in voice rehabilitation," *Auris Nasus Larynx*, vol. 34, pp. 327–332, Sep. 2007.

[5] G. S. Meltzner and R. E. Hillman, "Impact of aberrant acoustic properties on the perception of sound quality in electrolarynx speech," *J. Speech Lang. Hearing Res.*, vol. 48, pp. 766–779, Aug. 2005.

[6] C. Y. Espy-Wilson, V. R. Chari, J. M. Macauslan, C. B. Huang, and M. J. Walsh, "Enhancement of electrolaryngeal speech by adaptive filtering," *J. Speech Lang. Hearing Res.*, vol. 41, pp. 1253–1264, 1998.

[7] 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, pp. 670–678, 2003.

[8] L. Sheng, M. X. Wan, and S. P. Wang, "Multi-band spectral subtraction method for electrolarynx speech enhancement," *Algorithms*, vol. 2, pp. 550–564, 2009.

[9] R. L. Norton and R. S. Bernstein, "Improved laboratory prototype electrolarynx (LAPEL): Using inverse filtering of the frequency response function of the human throat," *Ann. Biomed. Eng.*, vol. 21, pp. 163–174, 1993.

[10] H. Liu, Q. Zhao, M. Wan, and S. Wang, "Application of spectral subtraction method on enhancement of electrolarynx speech," *J. Acoust. Soc. Amer.*, vol. 120, pp. 398–406, 2006.

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

[12] H. Gustafsson, S. E. Nordholm, and I. Claesson, "Spectral subtraction using reduced delay convolution and adaptive averaging," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 8, pp. 799–807, Nov. 2001.

[13] P. C. Pandey, S. M. Bhandarkar, G. K. Bachher, and P. K. Lehana, "Enhancement of alaryngeal speech using spectral subtraction," in *Proc. Int. Conf. Digit. Signal Process.*, 2002, vol. 2, pp. 591–594.

[14] F. Pernkopf, "Musical noise suppression for speech enhancement using pre-image iterations," in *Proc. 19th Int. Conf. Syst., Signals Image Process.*, 2012, vol. 22, pp. 464–467.

[15] T. Esch and P. Vary, "Efficient musical noise suppression for speech enhancement system," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2009, pp. 4409–4412.

[16] T. Hasan and M. K. Hasan, "Suppression of residual noise from speech signals using empirical mode decomposition," *IEEE Signal Process. Lett.*, vol. 16, no. 1, pp. 2–5, Jan. 2009.

[17] K. Paliwal, K. Jcicki, and B. Schwerin, "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain," *Speech Commun.*, vol. 52, pp. 450–475, 2010.

[18] S. Mudaliyar and N. Tahilramani, "Techniques of voice activity detection: A review," *Int. J. Sci. Res. Dev.*, vol. 5, pp. 1594–1597, 2017.
[19] T. S. Gunawan and M. Kartiwi, "On the characteristics of various quranic recitation for lossless audio coding application," in *Proc. Int. Conf. Comput. Commun. Eng.*, 2017, pp. 121–125.
[20] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *J. Acoust. Soc. Amer.*, vol. 95, pp. 1053–64, 1994.
[21] R. Drullman, J. M. Festen, and R. Plomp, "Effect of reducing slow temporal modulations on speech reception," *J. Acoust. Soc. Amer.*, vol. 95, pp. 2670–2680, 1994.
[22] T. Arai, M. Pavel, H. Hermansky, and C. Avendano, "Intelligibility of speech with filtered time trajectories of spectral envelopes," in *Proc. Int. Conf. Spoken Lang.*, vol. 4, 1996, pp. 2490–2493.
[23] H. Liu, M. Wan, and S. Wang, "Features of listeners affecting the perceptions of mandarin electrolaryngeal speech," in *Proc. Folia Phoniatrica Logopaedica Official Organ Int. Assoc. Logopedics Phoniatrics*, vol. 57, pp. 9–19, 2005.
[24] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL, USA: CRC Press, 2007.
[25] A. P. Pawar, K. B. Choudhari, and M. A. Joshi, "Review of single channel speech enhancement methods in spectral domain," *Int. J. Appl. Eng. Research*, vol. 7, pp. 1961–1966, 2012.
[26] R. M. Theodore, J. L. Miller, and D. Desteno, "Individual talker differences in voice-onset-time: Contextual influences," *J. Acoust. Soc. Amer.*, vol. 125, pp. 544–552, 2003.
[27] S. E. Blumstein, E. B. Myers, and J. Rissman, "The perception of voice onset time: An fMRI investigation of phonetic category structure," *J. Cogn. Neurosci.*, vol. 17, pp. 1353–1366, 2005.

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