An Improved Iterative Wiener Filtering Algorithm For Speech Enhancement

Ruitang Mao,Yi Zhou,Wenyi Yuan, and Hongqing Liu
School of Communication and Information Engineering
Chongqing University of Posts and Telecommunications, Chongqing, China
rt_mao@outlook.com, zhouy@cqupt.edu.cn, wy_y@outlook.com, hongqingliu@cqupt.edu.cn

***Abstract*—Speech enhancement is a critical technique for various applications like mobile communication and automatic speech recognition (ASR). This paper studies an improved iterative Wiener filtering (IWF) algorithm, which can help enhance the recognition rate of the ASR system. By using the voice activity detection (VAD) technology in the traditional IWF algorithm, the noise power spectrum estimation in the silent periods can be improved. Besides, the mini-tracking algorithm is also employed to estimate the signal-to-noise ratio (SNR) of the present speech signal, which is further used to control the number of iterations for the IWF algorithm. Therefore, the speech-like properties which are important in speech reconstruction and recognition are better retained. Computer simulations are conducted to verify the proposed algorithm has improved performance in ASR system over the conventional approach.**

Keywords—speech enhancement; VAD; Mini-tracking; IWF

I. INTRODUCTION

Speech enhancement is playing an increasingly important role to improve the performance of speech coding, automatic speech recognition (ASR), as well as many other applications in modern smart terminals. The ASR system does not always work in quiet environments and thus inevitably suffers from ambient noise and even competing speaker voice. These noise or interference can substantially lower the recognition rate. In order to solve this problem, remedy technology that removes noise from the noisy speech viz., speech enhancement, is strongly desired. An effective speech enhancement algorithm can raise the SNR of noisy speech and ensure the accuracy of endpoint detection in speech recognition preprocessing module. Over the past decades a lot of speech enhancement methods have been proposed [1]. Spectral subtraction (SS) algorithms [2] are relatively simple to implement, but are vulnerable to the so-called musical noise distortion. The Wiener filter (WF) algorithm [3] can reduce the musical noise. It was developed by Lim and Oppenheim based on minimizing mean squared error (MSE) criterion in the frequency domain. However, since the power spectrum of speech and additive noise cannot be obtained directly, WF algorithm has some practicality deficiency. To tackle this problem, proposed the iterative Wiener filter (IWF) algorithm which estimates the power spectrum of speech signal and noise respectively through the LPC analysis [4] of noisy speech and through processing the non-speech segments. The estimated noise power spectrum can thus be used to design the Wiener filter. Then, LPC analysis is applied to the enhanced speech and the Wiener filter is designed again. The above steps are repeated so as to get the more accurate speech power spectrum and to design more optimal Wiener filter. However, one significant deficiency of the IWF algorithm is the lack of proper convergence criteria, which will induce serious distortion to the estimated clean signal. For instance, if the optimal number of iteration cannot be determined, the speech formants will shift in location and decrease in formant bandwidth [5]-[6].

In this paper, from the aspects of ASR application, an improved IWF algorithm will be studied, which can help enhance the performance of the traditional IWF algorithm and the recognition accuracy of ASR system. The voice activity detection (VAD) technique based on the short-time energy can have a more accurate estimate of the noise power spectrum. Using it in the IWF algorithm can improve the estimate of the noise power spectrum in the silent periods. Especially, at low frequencies, due to the noise impact and the ambiguity of clean speech, the conventional IWF method has difficulty to correctly estimate the noise power spectrum. Besides, the VAD is also able to guarantee the endpoint detection in the preprocessing module of ASR system. On the other hand, we employ the mini-tracking algorithm [7] to estimate the signal-to-noise ratio (SNR) of the present speech signal, which is further employed to control the number of iterations so as to reduce computational cost and to prevent algorithm divergence. Hence, the proposed algorithm can better retain those speech-like properties such as MFCC and LPCC and improve the quality of speech reconstruction. Therefore, the recognition rate can also be increased.

The reminder of this paper is organized as follows. The traditional IWF algorithm is introduced in Section II. The VAD principle and the mini-tracking algorithm are analyzed in Section III. In Section IV, the improved IWF algorithm is derived. Computer experiments are conducted in Section V to evaluate the performance of the proposed algorithm. Finally, conclusions are drawn in Section VI.

II. ITERATIVE WIENER FILTER

In speech enhancement scenario, it is usually assumed that the noisy signal $y(n)$ can be written as

$$y(n) = x(n) + d(n), \tag{1}$$

where $x(n)$ is the clean speech signal, and $d(n)$ is the additive noise signal which is Gaussian distributed with zero

978-1-4673-8126-0 /15/$31.00 ©2015 IEEE

mean and variance σ_d^2. The IWF algorithm based on the autoregressive (AR) modeling [8] estimates the clean speech signal from noisy speech by iteratively applying the non-causal Wiener filter

$$H(w) = \frac{P_{xx}(w)}{P_{xx}(w) + P_{dd}(w)}. \quad (2)$$

In the case of eqns. (1)-(2), the filter can be defined as

$$X(w) = H(w)Y(w), \quad (3)$$

where w is the frequency index, $X(w)$, $Y(w)$, and $H(w)$ are the discrete Fourier transform of the clean speech, the noisy speech and the Wiener filter, respectively. $P_{dd}(w)$ and $P_{xx}(w)$ are the power spectral of $d(w)$ and $x(w)$. The $P_{xx}(w)$ in IWF algorithm is estimated as follows:

$$P_{xx}(w) = \frac{g^2}{|1 - \sum_{k=1}^{p} a_k e^{-jwk}|^2}, \quad (4)$$

where $\{a_k\}$ are the all-pole coefficients, which can be obtained by LPC analysis, and p is the number of all-pole coefficients. g is the gain of the system [9] which can be derived by

$$g^2 = \frac{\frac{2\pi}{N}\sum_{n=0}^{N-1} y^2(n) - 2\pi\sigma_d^2}{\int_{-\pi}^{\pi} \frac{1}{|1-\sum_{i=1}^{p} a_i e^{-jkw}|^2}}. \quad (5)$$

The performance of the IWF algorithm depends on the accuracy of power spectral estimation of the clean signal and additive noise. In the traditional IWF algorithm, $P_{dd}(w)$ is generally estimated in the first silent period, and $P_{xx}(w)$ is obtained through LPC analysis. In order to obtain a more clean speech, the iterative Wiener filtering process is repeated to obtain more clean speech. Due to the non-stationary nature of speech, this algorithm operates in short-time segments of the speech signal. In generally, a small number of iterations for convergent algorithm is necessary. However, because there is no suitable criterion can be used to determine when to stop the iterative algorithm, the best performance of IWF may not be achieved. To tackle this problem, We propose an improved method to control the convergence rate, which is crucial for improving the quality of speech reconstruction.

III. VAD AND MINI-TRACKING ALGORITHM

VAD is an indispensable technique used in speech enhancement. VAD can greatly eliminate the adverse effects of environmental background noise, which is helpful to improve the performance of voice application system. It distinguishes the active speech periods from silent pauses so as to ensure the accuracy of noise power spectrum estimation and endpoint detection, which are critical in ASR system [10]. Here, the efficient VAD scheme based on the energy threshold is employed to derive. Energy of a frame may represent possible presence to speech data, and voice data is an important parameter of the VAD algorithm.

$$E_j = \sum_{i=j}^{j+L-1} [y(i)w(j-i)]^2, \quad (6)$$

$$E_{\text{initial}} = \frac{1}{s}\sum_{k=1}^{s} E_j, \quad (7)$$

E_j = the noise energy of the j^{th} frame,

E_{initial} =initial threshold estimate,

$$|\tilde{D}(w)|^2_{initial} = \frac{\sum_{k=1}^{s}(|Y_k(w)|^2)}{s}. \quad (8)$$

Here, $|\tilde{D}(w)|^2_{\text{initial}}$ denotes the initial noise power spectrum. Assuming that the initial s sample does not contain any speech, L represents the frame length, and w is the window function. The initial threshold level for various parameters is computed from these samples. For example, the initial energy threshold is calculated from the mean of sample energy. Considering Eqs. (6)-(7), the final VAD decision can be made as [11]

$$t_{\text{VAD}}(j) = \begin{cases} 1, & \text{if } E_j > m\arg in * E_{\text{initial}} \\ 0, & \text{else} \end{cases} \quad (9)$$

where E_{initial} represents the initial threshold estimate, and the $m\arg in * E_j$ is the 'threshold' being used in the decision-making. The scaling factor $m\arg in$ allows a safe band for the adaptation of E_{initial}. Hence, When $t_{\text{VAD}}(j) = 0$, the noise power spectrum can be updated using the following the equation [12]

$$|\tilde{D}(w)|^2_{\text{new}} = G^*|\tilde{D}(w)|^2_{old} + (1-G)^*|Y_k(w)|^2, \quad (10)$$

where $|\tilde{D}(w)|^2_{\text{new}}$ denotes the updated noise power spectrum, and G is the smoothing factor.

The mini-tracking algorithm [13] was proposed for non-stationary noise environments. It makes use of the property that power level of the noisy speech signal lags behind the power level of the additive noise. So it can estimate the noise power spectrum by tracking the minimum value of the spectrum of the noisy speech signal. Its main principle can be briefly summarized as follows. By using a short sliding window, the minimum value of the spectrum within each

frequency band is found and compared with the power spectrum of the noisy speech after compensation. When the power spectrum of the noisy speech is less than the local minimum, the local minimum value needs to be updated. Meanwhile, in order to track and update the local minimum and the minimum value of the spectrum, the sliding window is further divided into sub-windows, and the estimated spectrum of the noise is updated in each sub-window. Therefore, the accuracy of the noise estimate can be improved. In this paper, the above method is used to estimate the current noise power spectrum in all active speech periods. However, the VAD scheme described above can only distinguish active from silent speech periods. Hence, in the very low SNR cases, it is difficult to detect the noise and unvoiced frames. With mini-tracking algorithm, the average of the SNR for all frames can be calculated. It can thus be used to set a classification criteria to stop the iterative algorithm.

$$\mathrm{SNR} = \frac{\sum_{i=0}^{L-1} |Y(w)|^2}{\sum_{i=0}^{L-1} |D(w)|^2}, \tag{11}$$

$$(\mathrm{SNR})_{\mathrm{mean}} = \frac{\sum_{n=0}^{N}\sum_{i=0}^{L-1} |Y(w)|^2}{\sum_{n=1}^{N}\sum_{i=0}^{L-1} |D(w)|^2}, \tag{12}$$

Here, N denotes the number of all frames of speech signal. $Y(w)$, and $D(w)$ represent the DFT of the corresponding signals. Based on Eqs. (11) and (12), the final decision on the number of iteration can be made as

$$Number_{\mathrm{iteration}} = \begin{cases} 2, & \text{if , SNR} > (\mathrm{SNR})_{\mathrm{mean}} + m\arg in \\ 1, & \text{else.} \end{cases} \tag{13}$$

IV. Improved IWF Algorithm

It is known that in speech enhancement algorithm based on IWF, the noise power spectrum of the current frame needs to be estimated. Usually, classical algorithms only employ the statistical average of silent periods to estimate the noise power spectrum. It is often assumed that the noise power outside of and within the vocal utterance be unchanged. Thus, this assumption makes the estimate obviously not comprehensive enough. In addition, the traditional IWF algorithm in practical applications has difficulty in determining the convergence condition so as to stop the iterative algorithm in time to ensure speech reconstruction quality. Therefore, in order to alleviate the above problems, an improved IWF algorithm is proposed in this section. On one hand, the noise power spectrum is estimated using speech endpoint detection algorithm instead of the traditional initial estimate method. In so doing, the deficiency of the classical algorithm in estimating the noise power spectrum in the silent period can be overcome. On the other hand, by using mini-tracking algorithm to estimate the SNR of the current speech signal, which is further used to control the number of iterations for the IWF algorithm. The quality of speech reconstruction and consequently the performance of the recognition system will be enhanced. Fig. 1 shows the block diagram of the improved IWF algorithm.

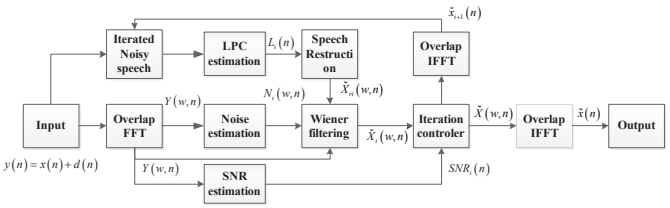


Fig. 1. Block diagram of the improved IWF algorithm

The improved IWF algorithm is summarized as follows:

For the each iteration:

Step 1: Using input signal and LPC analysis to obtain the parameters $a_0(n)$ of the initial all-pole model;

Step 2: Using FFT, through eqns. (6)-(10) to estimate the initial noise power spectrum $P_{dd}(w,n)$;

Step 3: Using $a_i(n)$, through eqns. (4)-(5) to reconstruct the estimate of clean speech $\tilde{P}_{x_i x_i}(w,n)$. Here, i represents the i^{th} iteration, where $i = 0,1,2$;

Step 4: With the results obtained in step 1 and 2, using eqns. (1)-(3) to design the optimal Wiener filter and obtain the estimated signal after speech enhancement $\tilde{X}_{i+1}(w,n)$;

Step 5: Using eqns. (11)-(13) to set the iteration controller;

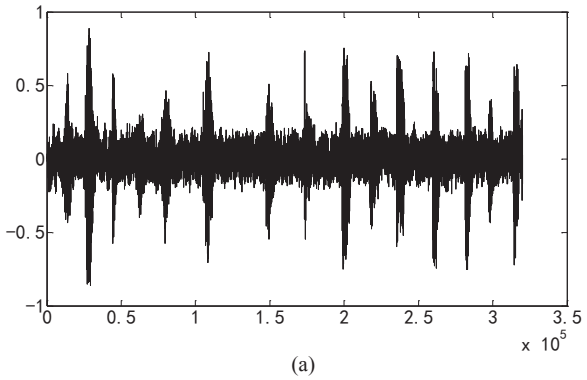
Step 6: With the results obtained in step 4, using IFFT and LPC analysis and returning to step 3 so that algorithm can continue the iteration until the iteration threshold is met;

Step 7: Finally, using the IFFT again to obtain the clean estimation signal $\tilde{x}_{i+1}(n)$ after speech enhancement.

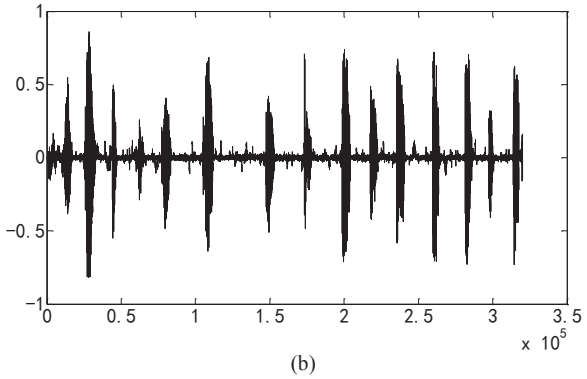
V. Simulation Results

In this section, the computer simulations are conducted to verify the efficiency of the proposed algorithm. Firstly, the traditional IWF algorithm, the spectral subtraction algorithm and the proposed IWF algorithm are tested for a comparison. The iteration number for the traditional IWF is fixed to 3. Pocket-sphinx [14], a speech recognition system developed by Carnegie Mellon University is employed to test the speech enhancement algorithms, with which the performances of all algorithms can be compared in the terms of recognition rate. The experimental speech data are extracted from a small thesaurus which contains 400 words sampled with 16 KHz rate. The additive noise data is the

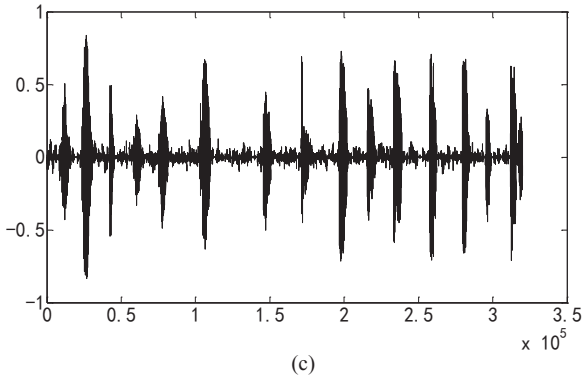
pink noise with various intensity of 2,4,6,…,30dB, respectively. The improved IWF algorithm is implemented as summarized by eqns. (1)-(13), where the relevant parameters are selected as: the smoothing factor $G = 0.9$; the safety parameter $margin = 0.01$; the initial iteration criterion $(\mathrm{SNR})_{\mathrm{mean}} = 23.78$. The initial iterative criterion can be obtained by calculating one of the above noisy speech data that is synthesized by a clean speech signal with a pink noise intensity of 10dB. The time-domain waveforms of speech signals before and after speech enhancement are shown in Fig. 2. The original noisy speech contains 30dB pink noise and lasts for 20 seconds.



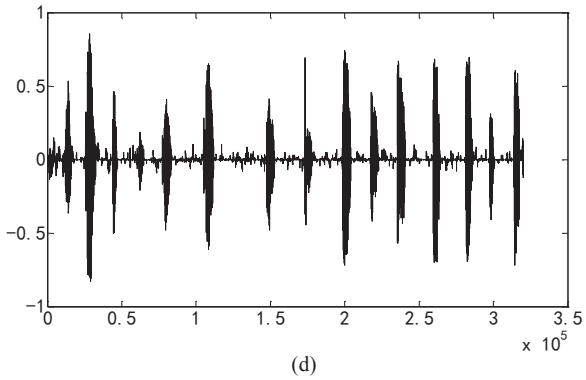
(a)



(b)



(c)



(d)

Fig. 2. (a) the original noisy speech, and the SE results using (b) the conventional SS algorithm, (c) the traditional IWF algorithm, (d) the improved IWF experiment algorithm.

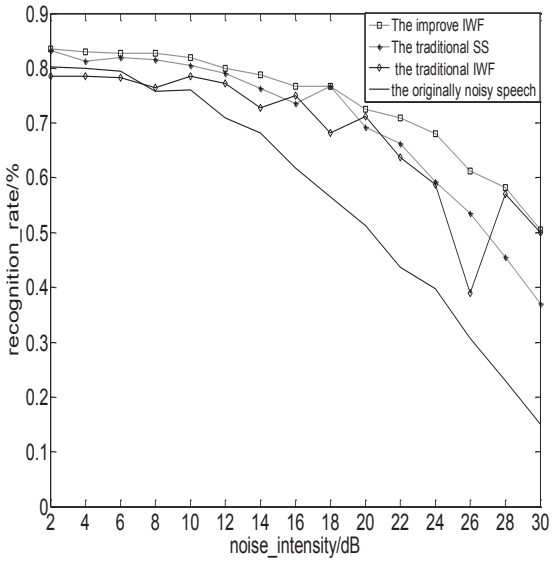


Fig. 3. Recognition rate vs. noise intensity for the SS, traditional IWF, and the proposed IWF algorithms.

As shown in Fig. 2, SS algorithm cannot get noise spectrum directly when reducing noise, which might lead to a great deviation between the estimated and true noise spectrum. Residual music noise with large fluctuations are generated. The iterative times for the traditional IWF algorithm is usually fixed which, for different kinds of speeches, will affect the quality of synthesized speech. Moreover, due to the parameters of the speech model extracted from noisy speech are inaccurate, the quality of synthesized speech will also be lowered. Given these shortcomings, the traditional IWF filtering algorithm is still better than the SS in reducing the musical noise. On the other hand, due to the inaccurate noise estimate and the lack of suitable convergence criteria, the speech formants are shifted in location and decreased in formant bandwidth. Therefore its actual noise reduction performance is inferior to the SS algorithm. In contrast, the improved IWF algorithm is comparatively more efficient than the

traditional IWF algorithm due to the controlled iterative times as well as the noise energy estimate. Consequently, the speech synthesized by LPC technique is of higher quality than the traditional IWF algorithm.

The recognition rates of the ASR system on noisy speeches before and after using the above speech enhancement algorithms are illustrated in Fig. 3. It can be shown that with the increase of noise intensity, all algorithms exhibit decreasing recognition rates. However, they all significantly improve the recognition rates after been applied to the original noisy signal. Due to the presence of music noise, the recognition rate of using SS algorithm does not strictly exponentially decrease. Also, it can be clearly seen that of all the compared algorithms, the proposed improved IWF algorithm has achieved the best performance, gaining the highest recognition rates at all SNR levels.

VI. CONCLUSIONS

In this paper, we proposed an improved iterative Wiener filtering algorithm which can improve the recognition rate of the automatic speech recognition system. The methodology employed includes the use of voice activity detection technology instead of the direct noise estimation method to effectively estimate the noise power spectrum in the traditional IWF, and the use of mini-tracking algorithm to estimate the signal-to-noise ratio (SNR) of the present speech signal, which is further employed to control the number of iterations. Consequently, the parameters representing speech-like properties are better retained. The speech recognition rate can thus be improved. Simulation results prove the advantage achieved by the proposed algorithm.

ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China (No. 61501072), Research Projects of Chongqing Educational Commission (KJ130504) and of Chongqing Science and Technology Commission (cstc2014jcyjA40017, cstc2015jcyjA40027).

REFERENCES

[1] M. Kato,et.al., “Noise suppression with high speech quality based on weighted noise estimation and MMSE STSA,” IEICE Trans. vol.E85-A. no.7, July 2002.

[2] S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” IEEE Trans., ASSP-27, pp. 113-120, 1979.

[3] J. S. Lim and A. V. Oppenheim, “All-pole modeling of degraded speech,” IEEE Trans., ASSP-26, pp.197-210, 1978.

[4] T. Shimamura and S. Takahashi, “Complex linear prediction method based on positive frequency domain,” IEICE Trans., vol. J72-A, pp. 1755-1763, 1989.

[5] J. H. L. Hansen and M. A. Clements, “Constrained iterative speech enhancement with application to speech recognition,” IEEE Trans. Signal Processing, vol. 39, pp. 795–805, 1991.

[6] T. V. Sreenivas and P. Kirnapure, “Codebook constrained Wiener filtering for speech enhancement,” IEEE Trans. Speech and Audio processing, vol. 4, pp. 383-389, September 1996.

[7] R. Martin, Spectral subtraction based on minimum statistics, Proceedings of the Seventh European Signal Processing Conference, EUSIPCO-94, Edinburgh, Scotland, 13–16 September 1994, pp. 1182-1185.

[8] E. Masgrau, J. Salavedra, A. Moreno, and A. Ardanuy, “Speech enhancement by adaptive Wiener filtering based on cumulant AR modeling,” Speech Processing in Adverse Conditions, M. Grenie and J. C. Junqua, Eds., 1992, pp. 143–146.

[9] P. C. Loizou, “Speech Enhancement, Theory and Practice,” CRC Press, 2007.9.

[10] T. Fukuda, O. Ichikawa, and M. Nishimura, “Long-term spectro-temporal and static harmonic features for voice activity detection,” Selected Topics in Signal Processing, IEEE Journal of, vol. 4, no. 5, pp. 834-844, 2010.

[11] B. Gold and N. Morgan, “Speech and audio signal processing,” John Wiley Publications.

[12] P. Pollak and P. Sovka, and J. Uhlir, “Noise suppression system for a car”, Proc. of the Third European Conference on Speech Communication and Technology -EUROSPEECH'93, Berlin, Sept 1993, pp 1073-1076.

[13] R. Martin, “An eDcient algorithm to estimate the instantaneous SNR of speech signals,” Proceedings of the Second European Conference on Speech, Communication and Technology, EUROSPEECH’93, Berlin, Germany, 21-23 September 1993, pp. 1093-1096.

[14] http://cmusphinx.sourceforge.net/wiki/download.