A Fast Convergence Speech Enhancement Method

Bing-yin XIA*, Chang-chun BAO*, Yan LIANG*

*Speech and Audio Signal Processing Lab, School of Electronic Information and Control Engineering
Beijing University of Technology, Beijing 100124, China
E-mail: xby-abc@emails.bjut.edu.cn, baochch@bjut.edu.cn, longmeimei@emails.bjut.edu.cn

Abstract—A fast convergence speech enhancement method is proposed in this paper. The noise estimation acceleration technique is applied to the conventional statistical model based algorithm to shorten the convergence time after the sudden change of noise intensity. First, the burst detection of power spectrum is performed on the noisy spectrum. Next, the log-likelihood ratio (LLR) based VAD is used in the period when the noise power is stationary, and the spectral entropy based VAD is implemented in the hang-over frames after the burst of noisy spectrum. Then a flag is set to control the update of noise estimation. Finally, an attenuated version of the noisy spectrum will be used directly as noise estimation if the update flag is set to one. The performance of the proposed method is evaluated under ITU-T G.160. In comparison with the conventional method, the convergence time is reduced evidently, while the abilities of noise reduction and SNR improvement are preserved, and the impact on the objective speech quality is constrained to a low level.

I. INTRODUCTION

The type and intensity of the background noise are always changing in the application of mobile communication. When a sudden change of noise power occurs, the residual noise in the enhanced speech will take a long time to converge to the specified level. The perceptual annoying residual noise in the convergence period will make speech quality degraded. So it is necessary to develop speech enhancement algorithms with the ability of fast convergence.

The key factor that determines the convergence time of speech enhancement is the tracking ability of noise estimation. There are many well known noise estimation methods, including the minimum statistics (MS) [1], minima controlled recursive averaging (MCRA) [2] and so on. Most of these methods are derived from the idea of tracking the minimum value of the noisy power spectrum. The minimum value is used either to estimate noise intensity directly or control the update speed of noise estimation. Usually, the minimum search is performed over a window containing D consecutive frames. When a sudden increase of the noise power happens, the minimum search will lag by 2D frames in the worst case scenario. As a result, the tracking delay of the noise estimation is between D and 2D frames. The typical value of window length is about 0.8-1.4 seconds. Then the convergence time might be as long as 3 seconds. This will make it not suitable for the performance requirement in the complex environment.

In this paper, a fast convergence speech enhancement method based on the burst detection of power spectrum and multi-parameter voice activity detection (VAD) is proposed. The convergence time is reduced evidently, while the amounts of noise reduction and signal-to-noise ratio (SNR) improvement are maintained, and the objective quality of enhanced speech will not be degraded beyond the acceptable range.

This paper is organized as follows. The details of the proposed method are described in Section II. The result of performance evaluation is given in Section III. And Section IV presents our conclusion.

II. THE PROPOSED ALGORITHM

In this paper, a fast convergence speech enhancement algorithm is proposed based on the noise estimation acceleration technique. The block diagram of the proposed algorithm is illustrated in Fig. 1. The noise estimation acceleration is applied to the traditional statistical model based weighted Euclidean distortion measure (WEDM) short-time spectral amplitude (STSA) estimator [3]. As shown in Fig. 2, the noise estimation acceleration consists of the burst detection of power spectrum, multi-parameter VAD and the update decision for noise estimation. If a sudden increase of the power spectrum is detected, an update flag will be set according to the VAD results of the hang-over frames. And the noise estimation will be updated if the flag is set to one. The details of the proposed method will be described in the following sub-sections.

A. The Burst Detection of Power Spectrum

The burst detection is intended to detect the abrupt increase of the noisy power spectrum, which may be a result of either the change of noise intensity or speech onset. The burst detection is carried out by the following steps:

Step 1: Calculate the ratio of the smoothed power spectrum to its long term average in each frequency bin;

Step 2: If the ratio is greater than a predefined threshold $T_{\text{inc}}=2\text{dB}$, the detection counter will be added by one;

![Fig. 1 Block diagram of the proposed method](image-url)
B. Multi-Parameter VAD Algorithm

In order to get high detection accuracy in the complex environment, a multi-parameter VAD algorithm is proposed in this paper. When the noise power is relatively stationary, the log-likelihood ratio (LLR) based VAD method could achieve satisfactory performance. While after the step change of noise intensity, the noise estimation will take a long time to converge. During this period, the spectral entropy based VAD method that is independent of noise estimation could be used. The combination of different VAD methods will perform better in the environment with non-stationary noise power.

The block diagram of the VAD algorithm is shown in Fig. 3.

First the burst detection of power spectrum is performed on the noisy speech. If the sudden change happens, the spectral entropy based VAD will be used in the hang-over frames. Otherwise, the method based on the LLR will be adopted.

a) Burst Detection for VAD algorithm

The burst detection in the VAD algorithm is carried out by the similar steps as described above. There are two main differences between the two methods. First, a higher threshold for the detection counter is used which will make the decision stricter. Second, the spectral deviation factor is used as an assistant parameter in the decision process.

Divide the spectrum into $M=32$ sub-bands with the same bandwidth and the spectral probability can be expressed as:

$$p(\lambda, i) = \frac{\sum_{k=1}^{N} |Y(\lambda, k)|^2}{\sum_{k=1}^{N} |Y(\lambda, k)|^2} \quad i=1,2,\ldots,M$$

(1)

where $p(\lambda, i)$ is the spectral probability of the $i^{th}$ subband in frame $\lambda$. $Y(\lambda, k)$ is the discrete Fourier transform (DFT) coefficient of the noisy speech at the $k^{th}$ bin of frame $\lambda$. $N=256$ is the length of DFT. $b_i$ and $e_i$ are the beginning and ending frequency bins of the $i^{th}$ subband, respectively.

The spectral deviation factor is defined as:

$$E(\lambda) = \frac{1}{i} \left( p(\lambda, i) - \bar{p}(i) \right)^2$$

(2)

where $\bar{p}(i)$ is the average spectral probability of the $i^{th}$ subband in the previous 5 frames.

For the stationary noise, the spectral deviation factor is relatively small, while for the speech onset, the value is much larger. Then if the value of the spectral deviation factor is greater than the threshold $T_{speech}=0.1$, the step change can be considered as a result of the speech onset. Otherwise, it may probably be caused by the burst of noise power.

b) VAD Algorithm based on LLR

The VAD algorithm based on LLR [4] is derived from the statistical model of speech signal. Assume that a clean speech signal $x(n)$ is corrupted by an additional noise signal $d(n)$. Taking Fourier transform of the noisy speech, we can get:

$$Y(\lambda, k) = X(\lambda, k) + D(\lambda, k)$$

(3)

where $Y(\lambda, k)$, $X(\lambda, k)$ and $D(\lambda, k)$ are the DFT coefficients of noisy speech, clean speech and noise signal, respectively.

For each frame, the two hypotheses for a VAD to consider are $H_1$ and $H_0$, which correspond to speech present and absent, respectively. Based on the Gaussian statistical model, the likelihood ratio of the $k^{th}$ frequency bin can be expressed as:

$$\Lambda_k = \frac{p(Y_k | H_1)}{p(Y_k | H_0)} = \frac{1}{1 + \xi_k^2} \exp \left\{ \frac{\gamma_k \delta_k}{1 + \xi_k^2} \right\}$$

(4)

where $p(Y_k | H_1)$ and $p(Y_k | H_0)$ are the probability density functions conditioned on $H_1$ and $H_0$, respectively. $\xi_k$ and $\gamma_k$ are the a priori and a posteriori SNR’s of frequency bin $k$, respectively. The estimations of $\xi_k$ and $\gamma_k$ can be got easily from the WEDM estimator.

The decision rule based on LLR is established using the geometric mean of the likelihood ratios for the individual frequency bins, which is given by:

$$\log \Lambda(\lambda) = \frac{1}{N} \sum_{k=1}^{N} \log \Lambda_k > \eta$$

(5)

where $\eta=2$ is the threshold for the VAD based on LLR.

If the value of LLR is greater than $\eta$, the current frame will be considered as speech, or it will be classified as noise.
c) VAD Algorithm based on Spectral Entropy

For the background noise whose spectral entropy distribution differs greatly from the one of speech signal, the spectral entropy based VAD could achieve relatively high decision accuracy. And it is independent of both the historical values and the noise estimation. Due to these advantages, this kind of method is adopted in the hang-over frames after the burst of noisy power spectrum.

The band-partitioning spectral entropy (BSE) [5] is used in this paper, which is defined as follows:

\[ H(\lambda) = -\sum_{i=1}^{\lambda} p(\lambda,i)\log p(\lambda,i) \]  

(6)

where \( p(\lambda,i) \) is the spectral probability defined above.

A first order recursive averaging is performed in the time domain to smooth the spectral entropy:

\[ \overline{H}(\lambda) = \alpha_{\text{entro}} \overline{H}(\lambda-1) + (1 - \alpha_{\text{entro}}) H(\lambda) \]  

(7)

where \( \alpha_{\text{entro}} = 0.8 \) is the smoothing factor.

The statistical distributions of the BSE for different kinds of signal are estimated and illustrated in Fig. 4.

![Fig. 4 BSE distribution histogram of noise and speech](image)

From the characteristics of BSE distribution, the noise signal can be classified into three categories.

For Type I, the value of BSE is smaller than that of the speech in most cases. The BSE distribution of Type II is in the right of the one for speech. While most part of the BSE histogram of Type III is overlapped with the speech histogram.

For the noise of Type I and II, the BSE is an appropriate feature to distinguish speech from noise. But the error probabilities of the BSE based VAD will increase for the noise of Type III.

During the process of speech enhancement, the noise BSE histogram is updated in the frames that are classified as speech absent by the LLR based VAD. Using the estimated noise BSE histogram and the speech histogram got from the training process, the background noise can be classified into different types, and different VAD schemes will be selected.

When the histogram estimation is not completed or the noise is classified as Type III, the method based on the search of extreme value will be adopted. Whether the maximum or the minimum value should be used is determined by the BSE parameters in the initial frames. The extreme value of BSE is searched using the similar method as the MS algorithm, and considered as an estimate of noise BSE parameter. Calculate the absolute difference between the BSE and its extreme value, if it is larger than a threshold \( T_s = 0.35 \), this frame will be classified as speech, or it will be considered as noise.

The histogram based method which is derived from the one in [6] is adopted for the noise of Type I and II. Based on the estimated noise BSE histogram and the trained one for speech, the threshold can be obtained using the principle of minimum error Bayesian decision. If the error probability for the speech is too high, the threshold will be modified by a constant value to ensure proper decision accuracy for the speech.

C. Noise Estimation Update Decision

Whether the noise estimation should be updated is decided according to the result of burst detection and the VAD flags in the hang-over frames.

There are three parameters in the process of update decision: \( K_0 \) is the number of speech absent frames in the hang-over period with a length of \( L_{\text{radist}} = 10 \) frames. \( E_c \) is the cumulative spectral deviation factor of the last \( L_{\text{dev}} = 5 \) frames:

\[ E_c = \sum_{i=1}^{\text{last}} E(i) \]  

(8)

\( T_{\text{err}} \) is the upper bound for the number of noise frames that are wrongly classified as speech, which can be expressed as:

\[ T_{\text{err}} = \left[ P_{\text{err}} \cdot L_{\text{radist}} + 0.5 \right] \]  

(9)

where \( P_{\text{err}} \) is the error probability of VAD for the noise, and \([x]\) is the operation of taking the integer part of \( x \).

If \( K_0 \) satisfies the following condition:

\[ K_0 \geq L_{\text{radist}} - T_{\text{err}} \]  

(10)

and \( E_c \) is smaller than the threshold \( T_{\text{err}} \), then the burst in the power spectrum is caused by the increase of the noise power, and the update of noise estimation is necessary.

If the update flag is set, the noisy power spectrum is attenuated to some extent and directly used as the noise estimation. As a result, the residual noise will be suppressed in a much shorter convergence time.

III. PERFORMANCE EVALUATION

The test of the proposed algorithm is carried out under the standard of ITU-T G.160 [7]. The main purpose is to evaluate the performance of speech enhancement in the amount of noise reduction and SNR improvement, the convergence time, and the objective quality of enhanced speech.

The clean speech sequences are chosen from the NTT database. The additive noises are taken from ITU-T and NoiseX-92 noise database, and some real-world noises are also used in our tests. Noise and speech sequences have been down-sampled to 8 kHz before the tests.

The performance of the proposed method is compared with the original WEDM STSA estimator in this section.

A. Noise Reduction and SNR Improvement Test

This test is intended to ensure that a speech enhancement method can produce expected amount of noise reduction and does not modify the speech level beyond the acceptable range. The test is carried out in white and colored noise, respectively.

The parameter \( Q_n \) is the specified level of noise reduction. For the test in white noise, \( Q_{n1} \) and \( Q_{n2} \) are the noise reduction factors in the periods that only contain noise. \( Q \) is the level difference of speech before and after speech enhancement.

74
For the test in colored noise, SNRI (Signal-to-Noise Ratio Improvement), TNLR (Total Noise Level Reduction) and DSN (SNRI to NPLR Difference, NPLR: Noise Power Level Reduction) are used to measure the SNR improvement and noise reduction abilities of the speech enhancement algorithm. The test result is presented in Table I.

### TABLE I

<table>
<thead>
<tr>
<th>Test Condition</th>
<th>Performance Parameters (dB)</th>
<th>WEDM</th>
<th>Proposed Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>White Noise</td>
<td>Q0</td>
<td>20.54</td>
<td>20.54</td>
</tr>
<tr>
<td></td>
<td>Q1</td>
<td>20.70</td>
<td>20.70</td>
</tr>
<tr>
<td></td>
<td>Q2</td>
<td>20.71</td>
<td>20.71</td>
</tr>
<tr>
<td>Colored Noise</td>
<td>Q3</td>
<td>0.45</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>SNRI</td>
<td>11.24</td>
<td>11.08</td>
</tr>
<tr>
<td></td>
<td>TNLR</td>
<td>-17.76</td>
<td>-17.78</td>
</tr>
<tr>
<td></td>
<td>DSN</td>
<td>1.00</td>
<td>1.31</td>
</tr>
</tbody>
</table>

It’s obvious that the introduction of the noise estimation acceleration technique will not impact the amounts of noise reduction and SNR improvement significantly in either kind of noise conditions, while the additional attenuation of speech component in colored noise remains low.

### B. Convergence Test for Noise Reduction

The convergence test is designed to ensure that a speech enhancement algorithm could produce specified level of noise reduction in response to the step change of noise intensity after a maximum allowed convergence time.

The convergence time is defined as the time from the sudden change of noise intensity to the instant when the level after a maximum allowed convergence time.

The test result is shown in Table II.

### TABLE II

<table>
<thead>
<tr>
<th>Noise Type</th>
<th>Test Result of Convergence Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Babble</td>
<td>WEDM</td>
</tr>
<tr>
<td></td>
<td>1.63</td>
</tr>
<tr>
<td>Office</td>
<td>0.00</td>
</tr>
<tr>
<td>Factory</td>
<td>0.00</td>
</tr>
<tr>
<td>Bus</td>
<td>0.00</td>
</tr>
<tr>
<td>Water</td>
<td>0.00</td>
</tr>
</tbody>
</table>

By adopting the proposed method, the convergence time, especially for $t_3$ which corresponds to the increase of noise power, can be reduced evidently for most of the noise types.

### C. Speech Quality Test for Noise Reduction

This test is meant to assess the perceptual quality of a speech enhancement algorithm. The test method is not described in the standard of ITU-T G.160. In this paper, PESQ is used as the objective measure of the speech quality. The test result is summarized in Table III.

### REFERENCES