Journal of Signal and Information Processing, 2013, 4, 314-326
http://dx.doi.org/10.4236/jsip.2013.43040 Published Online August 2013 (http://www.scirp.org/journal/jsip)

Single-Channel Speech Enhancement Using Critical-Band Rate Scale Based Improved Multi-Band Spectral Subtraction

Navneet Upadhyay[1], Abhijit Karmakar[2]

[1]Department of Electrical & Electronics Engineering, Birla Institute of Technology & Science, Pilani, India; [2]Integrated Circuit Design Group, CSIR—Central Electronics Engineering Research Institute, Pilani, India.
Email: navneet_upd@rediffmail.com, abhijit@ceeri.ernet.in

Received February 18th, 2013; revised March 18th, 2013; accepted April 18th, 2013

ABSTRACT

This paper addresses the problem of single-channel speech enhancement in the adverse environment. The critical-band rate scale based on improved multi-band spectral subtraction is investigated in this study for enhancement of single-channel speech. In this work, the whole speech spectrum is divided into different non-uniformly spaced frequency bands in accordance with the critical-band rate scale of the psycho-acoustic model and the spectral over-subtraction is carried-out separately in each band. In addition, for the estimation of the noise from each band, the adaptive noise estimation approach is used and does not require explicit speech silence detection. The noise is estimated and updated by adaptively smoothing the noisy signal power in each band. The smoothing parameter is controlled by *a-posteriori* signal-to-noise ratio (SNR). For the performance analysis of the proposed algorithm, the objective measures, such as, SNR, segmental SNR, and perceptual evaluations of the speech quality are conducted for the variety of noises at different levels of SNRs. The speech spectrogram and objective evaluations of the proposed algorithm are compared with other standard speech enhancement algorithms and proved that the musical structure of the remnant noise and background noise is better suppressed by the proposed algorithm.

Keywords: Single-Channel; Speech Enhancement; Critical-Band Rate Scale; Spectral Over-Subtraction; Adaptive Noise Estimation; Objective Measure; Speech Spectrograms

1. Introduction

Speech is the very basic way for humans to convey information to one another [1]. In many circumstances, the speech signals are severely degraded due to different types of background noises that limit their effectiveness for communication and make the listening difficult for a direct listener [2]. Therefore, the removal of noise components from the degraded speech and, in turn, its enhancement has been the main purpose of researches in the field of speech signal processing over the preceding few decades and it still remains as an open problem. The aim of the speech enhancement research is to minimize the effect of noises and make the speech more pleasant and understandable to the listener and thereby, to improve one or more perceptual aspects of speech, such as overall speech quality and/or intelligibility [3]. These two features, quality and intelligibility, are however uncorrelated to each other in a certain context. For example, a very clean speech of a speaker in a foreign language may be of high quality to a listener but at the same time it will be of zero intelligibility. Therefore, a low quality speech may be high in intelligibility while a high quality speech may be low in intelligibility [4].

The classification of speech enhancement methods depends on the number of microphones that are used for recording speech data into single, dual or multi-channel. Though the performance of multi-channel speech enhancement is found to be better than single-channel speech enhancement [1,3], the single-channel speech enhancement is still a significant field of research interest because of its simple implementation. Single-channel speech enhancement method requires the noise estimation process during speech silences. The estimation of the spectral magnitude from the noisy speech is easier than

the estimate of both magnitudes and phases. In [5], it is revealed that the short-time spectral magnitude (STSM) is more important than phase information for intelligibility and quality of speech signals.

The spectral subtraction method proposed by Boll [6] is one of the widely used single-channel speech enhancement approaches based on the direct estimation of STSM. The main attraction of the spectral subtraction method is: 1) its relative simplicity. It only requires an estimate of the noise spectrum; and 2) its high flexibility against subtraction parameters variation. Usually, the spectral subtraction method uses the statistical information of silence region, detected by a voice activity detector (VAD). However, if the background noise is non-stationary, it will be difficult to use VAD. Also, the enhanced speech obtained by conventional spectral subtraction has shortcomings that the speech contains the perceptually noticeable spectral artifacts, known as the remnant musical noise, which is composed of unnatural artifacts with random frequencies and perceptually annoys the human ear. In recent years, a number of speech enhancement algorithms have been proposed which deal with the modifications of the spectral subtraction method to combat the problem of remnant musical noise artifacts [7-11] and improve the quality of speech in noisy environments.

In this paper, critical-band rate scale based on improved multi-band spectral subtraction algorithm is proposed for enhancement of the single-channel speech. The proposed algorithm (PM) uses a new noise estimation approach to estimate the noise adaptively and continuously from the nearby previous speech frames without explicit speech silence detection. In addition to this, a smoothing parameter is used which is controlled by *a-posteriori* SNR. The PM attempts to find the optimal trade-off between speech distortion and noise reduction.

The paper is organized as follows: In Section 2, we describe the principle of the spectral subtraction method [6], spectral over-subtraction algorithm [7] for enhancement of degraded speech and a noise estimation approach is presented to estimate the noise. In Section 3, we develop and present critical-band rate scale based on improved multi-band spectral subtraction algorithm for enhancement of speech in adverse conditions. Experimental results and performance evaluations are done in Section 4, followed by the conclusions in Section 5.

2. Principle of Spectral Subtraction Method

In real-world conditions, the speech signal is generally degraded by additive noise [3,6]. Additive noise is typically the background noise and is uncorrelated with the clean speech signal. The speech signal degraded by background noise is named as noisy speech. The noisy signal can be modeled as the sum of the clean speech signal and the random noise [3,6] as

$$y(n) = s(n) + d(n),\ n \in (0, N-1) \tag{1}$$

where n is the discrete time index and N is the number of samples in the signal. Here, $y(n)$, $s(n)$, and $d(n)$ are the n^{th} sample of the discrete time signal of noisy speech, clean speech and the noise, respectively. As the speech signal is non-stationary in nature and contains transient components, usually the short-time Fourier transform (STFT) is used to divide the speech signal in small frames for further processing, in order to make it stationary over the frames. Now representing the STFT of the time windowed signals by $Y_W(\omega)$, $D_W(\omega)$, and $S_W(\omega)$, (1) can be written as [6,12],

$$Y_W(\omega) = S_W(\omega) + D_W(\omega) \tag{2}$$

where ω is the discrete frequency index of the frame.

The spectral subtraction method mainly involves two stages. In the first stage, an average estimate of the noise spectrum is subtracted from the instantaneous spectrum of the noisy speech. This is named as basic spectral subtraction step. In the second stage, several modifications like half-wave rectification (HWR), remnant noise reduction and signal attenuation are done to reduce the signal level in the non-speech regions. In the entire process, the phase of noisy speech is kept unchanged because it is assumed that the phase distortion is not perceived by human ear [5]. Therefore, the STSM of noisy speech is equal to the sum of STSM of clean speech and STSM of noise without the information of phase and (2) can be expressed as

$$|Y_W(\omega)| = |S_W(\omega)| + |D_W(\omega)| \tag{3}$$

where $Y_w(\omega) = |Y_w(\omega)|\exp\left(j\varphi_y(\omega)\right)$,
$S_w(\omega) = |S_w(\omega)|\exp\left(j\varphi_y(\omega)\right)$,
$D_w(\omega) = |D_w(\omega)|\exp\left(j\varphi_y(\omega)\right)$ and $\varphi_y(\omega)$ is the phase of the noisy speech. To obtain the short-time power spectrum of noisy speech, $Y_w(\omega)$ is multiplied by its complex conjugate $Y_w^*(\omega)$. In doing so, (2) become

$$\begin{aligned}|Y_w(\omega)|^2 = |S_w(\omega)|^2 + |D_w(\omega)|^2 + S_w(\omega)D_w^*(\omega) \\ + S_w^*(\omega)D_w(\omega)\end{aligned} \tag{4}$$

Here $D_w^*(\omega)$ and $S_w^*(\omega)$ are the complex conjugates of $D_w(\omega)$ and $S_w(\omega)$ respectively. The terms $|Y_w(\omega)|^2$, $|S_w(\omega)|^2$, and $|D_w(\omega)|^2$, are referred to as

the short-time power spectrum of noisy speech, clean speech, and random noise respectively. In (4), the terms $\left|D_w(\omega)\right|^2$, $S_w(\omega)D_w^*(\omega)$ and $S_w^*(\omega)D_w(\omega)$ cannot be obtained directly and are approximated as, $E\left\{\left|D_w(\omega)\right|^2\right\}$, $E\left\{S_w(\omega)D_w^*(\omega)\right\}$ and $E\left\{S_w^*(\omega)D_w(\omega)\right\}$, where $E\{.\}$ denotes the ensemble averaging operator. As the additive noise is assumed to be zero mean and orthogonal with the clean speech, the terms $E\left\{S_w(\omega)D_w^*(\omega)\right\}$ and $E\left\{S_w^*(\omega)D_w(\omega)\right\}$ reduce to zero [3]. Therefore, (4) can be rewrite as

$$\begin{aligned}\left|\hat{S}_w(\omega)\right|^2 &= \left|Y_w(\omega)\right|^2 - E\left\{\left|D_w(\omega)\right|^2\right\} \\ &= \left|Y_w(\omega)\right|^2 - \left|\hat{D}_w(\omega)\right|^2\end{aligned} \tag{5}$$

where $\left|\hat{S}_w(\omega)\right|^2$ and $\left|Y_w(\omega)\right|^2$ is the short-term power spectrum of estimated speech and the noisy speech, respectively. The $\left|\hat{D}_w(\omega)\right|^2$ is the estimated noise power spectrum [12].

In spectral subtraction method, it is assumed that the speech signal is degraded by additive white Gaussian noise (AWGN) with flat spectrum; therefore, the noise affects the signal uniformly over the complete spectrum. In this method, the subtraction process needs to be done cautiously to avoid any speech distortion. The spectra obtained after subtraction process may contain some negative values due to inaccurate estimation of the noise spectrum. Since, the spectrum of estimated speech can become negative due to over-estimation of noise, but it cannot be negative, therefore a half-wave rectification (setting the negative portions to zero) or full-wave rectification (absolute value) are introduced. Half-wave rectification (HWR) is commonly used but it introduces annoying noise in the enhanced speech. Full-wave rectification (FWR) avoids the creation of annoying noise, but it is less effective in suppressing noise. Thus, the complete power spectral subtraction algorithm is given by (see Equation (6))

As the human perception is insensitive to phase [5], the enhanced speech is reconstructed by taking the inverse STFT (ISTFT) of the enhanced spectrum using the phase of the noisy speech and overlaps-add (OLA) method, can be expressed as

$$\hat{s}_w(n) = \text{ISTFT}\left\{\left|\hat{S}_w(\omega)\right|\exp\left(j\varphi_y(\omega)\right)\right\} \tag{7}$$

On the contrary, a generalized form of spectral subtraction method (6) can be obtained by changing the power exponent from 2 to b, which determines the sharpness of the transition.

$$\left|\hat{S}_w(\omega)\right|^b = \left|Y_w(\omega)\right|^b - \left|\hat{D}_w(\omega)\right|^b,\ b>0 \tag{8}$$

where $b=2$ represents the power spectrum subtraction and $b=1$ represents the magnitude spectrum subtraction.

The drawback of spectral subtraction method is that it suffers from some severe difficulties in the enhancement process. From (5), it is clear that the effectiveness of spectral subtraction is heavily dependent on accurate noise estimation, which additionally is limited by the performance of speech/pause detectors. A VAD performance degrades significantly at lower SNR. When the noise estimate is less than perfect, two major problems occur, remnant residual noise, referred as musical noise, and speech distortion. The spectral over-subtraction algorithm proposed by Berouti [7] is the improvement of magnitude spectral subtraction algorithm [6].

2.1. Spectral Over-Subtraction Algorithm

An improved version of spectral subtraction method was proposed in [7] to minimize the annoying noise and distortion. In this algorithm, the spectral subtraction method [6] uses two additional parameters, over-subtraction factor, and noise spectral floor parameter β [7]. The algorithm is given as (see Equation (9))
with $\alpha \geq 1$ and $0 \leq \beta \ll 1$.

The over-subtraction factor controls the amount of noise power spectrum subtracted from the noisy speech power spectrum in each frame and spectral floor parameter prevents the resultant spectrum from going below a preset minimum level rather than setting to zero (spectral floor). The over-subtraction factor depends on a-*pos*-

$$\left|\hat{S}_w(\omega)\right|^2 = \begin{cases}\left\{\left|Y_w(\omega)\right|^2 - \left|\hat{D}_w(\omega)\right|^2\right\}, & \text{if } \left|Y_w(\omega)\right|^2 > \left|\hat{D}_w(\omega)\right|^2 \\ 0 & \text{else}\end{cases} \tag{6}$$

$$\left|\hat{S}_w(\omega)\right|^2 = \begin{cases}\left|Y_w(\omega)\right|^2 - \alpha\cdot\left|\hat{D}_w(\omega)\right|^2, & \text{if } \dfrac{\left|\hat{D}_w(\omega)\right|^2}{\left|Y_w(\omega)\right|^2} < \dfrac{1}{\alpha+\beta} \\ \beta.\left|\hat{D}_w(\omega)\right|^2, & \text{else}\end{cases} \tag{9}$$

teriori segmental SNR. The over-subtraction factor can be calculated as

$$\alpha = \alpha_0 + (\mathrm{SNR})\left(\frac{\alpha_{\min} - \alpha_0}{\mathrm{SNR}_{\max}}\right) \quad (10)$$

Here

$\alpha_{\min} = 1, \alpha_{\max} = 5, \mathrm{SNR}_{\min} = -5$ dB, $\mathrm{SNR}_{\max} = 20$ dB and $\alpha_0\,(\alpha_0 \approx 4)$ is the desired value of α at 0 dB SNR . These values are estimated by experimental trade-off results. The relation between over-subtraction factor and SNR is shown in **Figure 1**.

This implementation assumes that the noise affects the speech spectrum uniformly and the subtraction factor subtracts an over-estimate of noise from noisy spectrum. Therefore, for a balance between speech distortion and remnant musical noise removal, various combinations of over-subtraction factor α, and spectral floor parameter β give rise to a trade-off between the amount of remnant noise and the level of perceived musical noise. For large value of parameter β, a very little amount of remnant musical noise is audible, while with small β, the remnant noise is greatly reduced, but the musical noise becomes quite annoying. Therefore, the suitable value of α is set as per (10) and $\beta = 0.03$.

This algorithm reduces the level of perceived remnant musical noise, but background noise remains present and enhanced speech is distorted.

2.2. Noise Estimation

In real-world environment, the noise does not affect the speech signal uniformly over the complete frequency spectrum. Some of the frequency components of speech are affected more adversely than others due to this type of noise. This kind of noise is referred as non-stationary or colored noise [12]. Therefore, the noise spectrum estimation is the fundamental requirement of speech enhancement algorithm. If the noise estimate is too low, annoying remnant noise will be audible, and if the noise estimate is too high, speech will be distorted, possibly resulting in intelligibility loss. There are many methods to estimate the noise power, especially during speech activity. The non-stationary noise power can be estimated using minimal-tracking algorithms [13], and time-recursive averaging algorithms [14,15]. In the recursive-averaging type of algorithms [14,15], the noise spectrum is estimated as a weighted average of the past noise estimates and the present noisy speech spectrum. The weights change adaptively depending on the effective SNR of each frequency bin. In this paper, the non-stationary noise estimate is updated by adaptively smoothing the noisy signal power as a sum of the past noise power and the present noisy signal power without the need of an explicit speech pause detection. The smoothing parameter is controlled by a linear function of *a-posteriori* SNR.

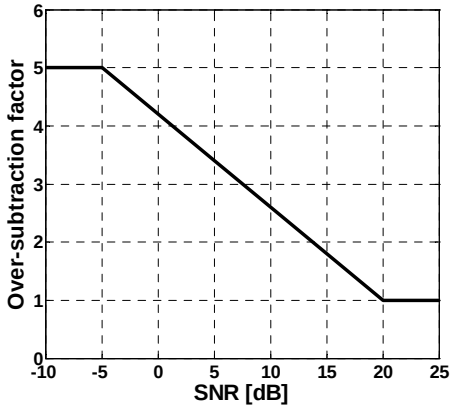


Figure 1. The relation between over-subtraction factor and SNR.

The noise estimate can be calculate as first order recursive algorithms as

$$\hat{\sigma}_d^2(\omega,k) = \lambda(\omega,k)\hat{\sigma}_d^2(\omega,k-1) + (1-\lambda(\omega,k))\left|Y(\omega,k)\right|^2 \quad (11)$$

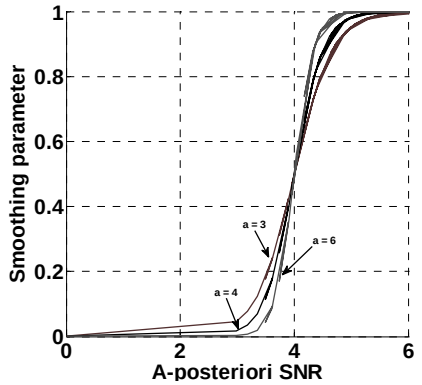
where k is the frame index, ω is the frequency bin index, $\sigma_d^2(\omega,k)$ is the noise power spectrum estimation in the ω^{th} frequency bin of current frame and $\left|Y(\omega,k)\right|^2$ is the short-time power spectrum of noisy speech. Further, $\lambda(\omega,k)$ is a time and frequency dependent smoothing parameter whose value depends on the noise changing rate.

The smoothing parameter is the time-varying frequency dependent parameter that is adjusted by the speech presence probability. In [16], the smoothing parameter $\lambda(\omega,k)$ at frame k is selected as a sigmoid function changing with the estimate of the *a-posteriori* $\mathrm{SNR}(k)$.

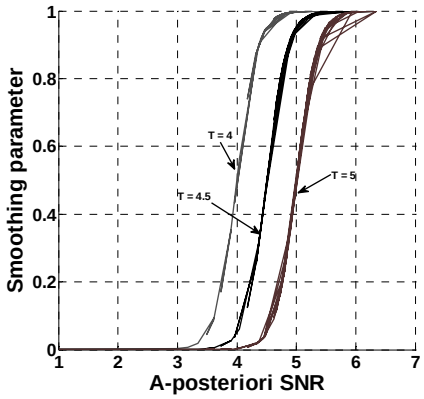
$$\lambda(\omega,k) = \frac{1}{1+\exp\left(-a\left(\mathrm{SNR}(\omega,k) - T\right)\right)} \quad (12)$$

where parameter a in sigmoid function (12) affects the noise changing rate and is a constant with a value between 1 to 6. The parameter T in (12) is the center-offset of the transition curve in sigmoid function and the value of T are around 3 to 5. A plot of smoothing parameter against the *a-posteriori* SNR at different values of a and different values of T is shown in **Figures 2(a)** and **(b)**, respectively.

The update of noise estimate must be performed only in the absence of speech at the corresponding frequency



(a)



(b)

Figure 2. Plot of smoothing parameter against the *a-posteriori* SNR: (a) for different value of α, and (b) for different values of *T*.

bin. This can be performed by controlling the smoothing factor $\lambda(\omega,k)$ depending on the *a-posteriori* $\mathrm{SNR}(k)$ defined as

$$\mathrm{SNR}(\omega,k)=10\log_{10}\left(\frac{|Y(\omega,k)|^2}{\frac{1}{m}\sigma_d^2\sum_{p=1}^{m}(\omega,k-p)}\right) \quad (13)$$

where the denominator part is the average of the noise estimate of the previous m frames (previous 5 to 10 frames) immediately before the frame k.

In [16], a different function was proposed for computing $\lambda(\omega,k)$ as

$$\lambda(\omega,k)=1-\min\left\{1,\frac{1}{(\mathrm{SNR}(\omega,k))^p}\right\} \quad (14)$$

where p is an integer, and $\mathrm{SNR}(\omega,k)$ is given by (13).

The slope parameter a in (12) controls the way in which smoothing parameter $\lambda(\omega,k)$ changes with *a-posteriori* SNR. Generally, larger values of α in (12) lead to larger values of $\lambda(\omega,k)$ and slower noise updates, whereas smaller values of a in (12) give faster noise updates, at the risk of possible over-estimation during long voiced intervals. It results in smoothing parameter being close to 0 when the speech is absent in frame k, that is, the estimate of noise power in frame k follows rapidly the power of the noisy signal in the absence of speech. On the other hand, if speech signal is present, the new noisy signal power is much larger than the previous noise estimate. Thus, the value of smoothing parameter increases rapidly with increasing SNR. Hence, the noise update is slower or eventually stops because of the larger value of the smoothing parameter. Theoretically, the *a-posteriori* SNR should always be 1 when noise alone is present and greater than 1 when both speech and noise are present.

The main advantage of using the time-varying smoothing factor $\lambda(\omega,k)$, is that the noise power will be adapted differently and at different rates in the various frequency bins, depending on the estimate of the *a-posteriori* $\mathrm{SNR}(k)$ in that bin.

3. Critical-Band Rate Scale Based Improved Multi-Band Spectral Subtraction Algorithm

It is well-known that the sensitivity of human ear varies non-linearly in the frequency spectrum [17]. Therefore, the notion of critical-band is important for describing hearing sensations such as perception of loudness, pitch, and timbre. A commonly used scale for signifying the critical-bands is the critical-band rate scale. The critical-band rate scale divides the range of human auditory frequency spreads from 20 Hz to 20 kHz into 24 critical-bands (CBs). However, the frequency bandwidth of the narrowband human voice is typically only 4 kHz. Therefore, the bands in the proposed algorithm are derived in such a manner that it closely matches the psychoacoustic frequency scale of human ear.

Based on the measurements by Zwicker *et al.* [18], the critical-band rate scale can approximately be expressed in terms of the linear frequency as

$$z(f)=13\tan^{-1}\left(7.6\times10^{-4}f\right)+3.5\tan^{-1}\left(1.33\times10^{-4}f\right)^2 \quad (15)$$

Here, $z(f)$ is the critical-band rate scale in Bark, and f is the frequency in Hz. For the implementation of our proposed algorithm, the sampling rate is chosen to be 8 kHz. In **Figure 3**, a mapping between the physical (linear) frequency scale and the critical-band rate scale is shown [18]. The corresponding critical bandwidth (CBW) of the center frequencies can be expressed by

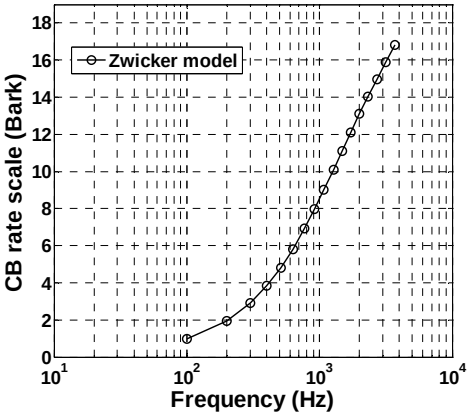


Figure 3. Mapping between the physical frequency scale and critical-band rate scale.

$$\mathrm{CBW}\left(f_c\right)=25+75\left(1+1.4\times10^{-6}f_c^2\right)^{0.69} \tag{16}$$

where f_c is the center frequency (Hz). Within this bandwidth, there are approximately 18 critical-bands as listed in **Table 1** [18]. According to the specifications of center frequencies, lower and upper edge frequencies of the CBs are given in **Table 1**.

In this paper, the linearly frequency spaced multi-band approach [8] is modified to non-uniformly frequency spaced bands [19]. Therefore, the estimate of the clean speech spectrum in the i^{th} non-uniformly spaced frequency band [20] is obtained by

$$\left|\hat{S}_i\left(\omega\right)\right|^2=\begin{cases}\left|Y_i\left(\omega\right)\right|^2-\alpha_i\cdot\delta_i\cdot\hat{\sigma}_{d_i}^2\left(\omega\right), \\ \text{if } \left|\hat{S}_i\left(\omega\right)\right|^2>\beta\cdot\left|Y_i\left(\omega\right)\right|^2 \\ \beta\cdot\left|Y_i\left(\omega\right)\right|^2 \qquad \text{else}\end{cases} \tag{17}$$

where $k_i<\omega<k_{i+1}$.

Here, k_i and k_{i+1} are the beginning and ending frequency edges of the i^{th} non-uniformly spaced frequency band. The α_i is the over-subtraction factor for the non-uniformly frequency spaced band-specific analysis, which is the function of the segmental SNR. The segmental SNR_i is computed using spectral components from each Band i as

$$\sum_{\omega=k_i}^{k_{i+1}}\left|Y_i\left(\omega\right)\right|^2,\sum_{\omega=k_i}^{k_{i+1}}\hat{\sigma}_{d_i}^2\left(\omega\right),\ i=1,2,\cdots,K \tag{18}$$

Here i is the band number, K is the total number of non-uniformly spaced frequency bands. The segmental SNR of the i^{th} Band can be calculated as

$$\mathrm{SNR}_i\left(\mathrm{dB}\right)=10\log_{10}\left(\frac{\sum_{\omega=k_i}^{k_{i+1}}\left|Y_i\left(\omega\right)\right|^2}{\sum_{\omega=k_i}^{k_{i+1}}\hat{\sigma}_{d_i}^2\left(\omega\right)}\right) \tag{19}$$

Here $\hat{\sigma}_{d_i}^2\left(\omega\right)$ is estimated using (11). The band specific over-subtraction can be calculated, using **Figure 1**, as (see Equation (20)).

The scale factor δ_i, in (17), is used to provide an additional degree of control over the noise subtraction level in each non-uniformly spaced frequency band. The values of δ_i is empirically determined and set to

$$\delta_i=\begin{cases}0.8, & f_i\le1\mathrm{kHz} \\ 1.3, & 1\mathrm{kHz}<f_i\le\dfrac{f_s}{2}-2\mathrm{kHz} \\ 1, & f_i>\dfrac{f_s}{2}-2\mathrm{kHz}\end{cases} \tag{21}$$

where f_i is the ending frequency of i^{th} Band and f_s is the sampling frequency. Since most of the speech energy is present in the lower frequencies, smaller values of δ_i are used for the low-frequency bands in order to minimize speech distortion.

It is computationally inefficient to separate the whole speech spectrum into such a large number of non-uniformly spaced frequency intervals, as given in **Table 1**, because it is very difficult to set the value of additional band over-subtraction empirically, for each CB, separately. Thus, the CBs, as in **Table 1**, are grouped together into six non-uniformly spaced frequency bands each containing three conjugative CBs. Therefore, the speech spectrum analysis is performed in a total number of six non-uniformly spaced frequency bands, closely matching with the non-uniform frequency spacing given by the human auditory system. In our algorithm, the speech spectrums has been divided into six frequency bands with ranges of {20 Hz - 300 Hz (Band 1), 300 Hz - 630 Hz (Band 2), 630 Hz - 1080 Hz (Band 3), 1080 Hz - 1720 Hz (Band 4), 1720 Hz - 2700 Hz (Band 5), 2700 Hz - 4 kHz (Band 6)}, and spectral over-subtraction is carried-out independently in these $K\left(K=6\right)$ non-overlapping non-uniformly spaced frequency bands.

$$\alpha_i=\begin{cases}\alpha_{\max}, & \text{if } \mathrm{SNR}_i\le\mathrm{SNR}_{\min} \\ \alpha_{\max}+\left(\mathrm{SNR}_i-\mathrm{SNR}_{\min}\right)\left(\dfrac{\alpha_{\min}-\alpha_{\max}}{\mathrm{SNR}_{\max}-\mathrm{SNR}_{\min}}\right), & \text{if } \mathrm{SNR}_{\min}\le\mathrm{SNR}_i\le\mathrm{SNR}_{\max} \\ \alpha_{\min}, & \text{if } \mathrm{SNR}_i\ge\mathrm{SNR}_{\max}\end{cases} \tag{20}$$

Table 1. Critical-bands rate scale of the human auditory system for frequency bandwidth of 4 kHz.

CB rate (Bark)	Lower edge freq. (Hz)	Upper edge freq. (Hz)	Center freq. (Hz)	CBW (Hz)
1	20	100	50	100
2	100	200	150	100
3	200	300	250	100
4	300	400	350	100
5	400	510	450	110
6	510	630	570	120
7	630	770	700	140
8	770	920	840	150
9	920	1080	1000	160
10	1080	1270	1170	190
11	1270	1480	1370	210
12	1480	1720	1600	240
13	1720	2000	1850	280
14	2000	2320	2150	320
15	2320	2700	2500	380
16	2700	3150	2900	450
17	3150	3700	3400	550
18	3700	4000	3850	-

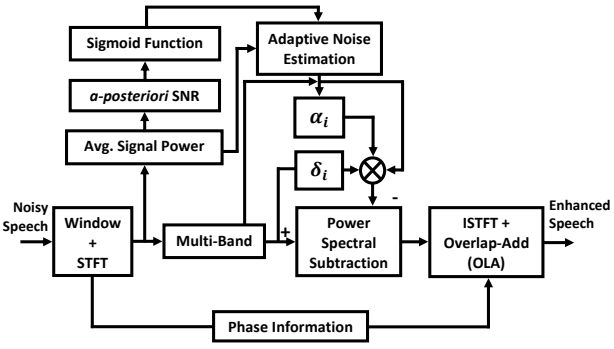


Figure 4. Block diagram of proposed algorithm.

In **Figure 4**, the block diagram of improved multi-band spectral subtraction algorithm based on critical-band rate scale for speech enhancement is shown.

4. Experimental Results and Performance Evaluation

This section presents the experimental results and performance evaluation of the proposed enhancement algorithm as well as comparison with the basic spectral subtraction (BSS) algorithm and multi-band spectral subtraction (MBSS) algorithm. For simulations, we have employed MATLAB software as the simulation environment. The noisy speech samples have been taken from NOIZEUS speech corpus [21]. The NOIZEUS is comprised of 30 phonetically balanced sentences belonging to six speakers, three male and three female, and degraded by seven different real-world noises at different levels of SNRs. The corpus is sampled at 8 kHz and quantized linearly using 16 bits resolution. A total of four different utterances, (three male speakers and one female speaker), from NOIZEUS corpus, are used in our evaluation. The noises have different time-frequency distributions, and therefore a different impact on speech. Hence, eight types of noises, seven real-world noise and a computer generated white Gaussian noise, have been used for the evaluation of the proposed speech enhancement algorithm. The real-world noises are car, train, restaurant, babble, airport, street, and exhibition. The performance of the proposed speech enhancement system is tested on such noisy speech samples.

In our experiments, the frame size is chosen to be 256 samples, *i.e.*, a time frame of 32 ms, with 50% overlapping. The sinusoidal Hamming window with size 256 samples is applied to the noisy signal. The noise estimate is updated adaptively and continuously using the smoothing parameter (11). For calculation of smoothing parameter, the value of a and T is chosen to be 4 and 5, respectively in the sigmoid function (12).

For the comparison purpose, two classes of multi-band spacing are employed in this paper. Firstly, we use a uniformly frequency spaced multi-band spectral subtraction algorithm where the over-subtraction factor α_i is computed for each uniformly spaced frequency band [8]. In this algorithm, four uniformly spaced frequency bands {60 Hz - 1 kHz (Band 1), 1 kHz - 2 kHz (Band 2), 2 kHz - 3 kHz (Band 3), 3 kHz - 4 kHz (Band 4)} have been taken. The value of over-subtraction factor α_i is determined using **Figure 1** and (20), and the value of additional over-subtraction factor δ_i for each band is set as per [8]. The value of spectral flooring parameter β is taken as 0.03 [8] and noise estimate is updated during the silence frames by using averaging.

For our proposed algorithm (PM), several implementations with various numbers of bands have been considered [19,20]. It has been found that the performance of the algorithm does not improve, for bands numbering more than six. Thus, the CBs, as in **Table 1**, are grouped together into six non-uniform bands and each band containing three conjugative CBs. Therefore, the spectrum analysis is performed in a total number of six non-uniformly spaced frequency bands, closely matching with the non-uniform frequency spacing given by the human au-

ditory system. These numbers of bands gives an optimal speech quality. The noise in each band is estimated by using the adaptive noise estimation approach as presented in Section II. The value of α_i is calculated as per (20) and δ_i is fixed as per (21). The value of spectral flooring parameter parameters has been taken to be same as the reference algorithm (MBSS) [8].

To test the performance of proposed speech enhancement algorithm, the objective quality measurement tests, signal-to-noise ratio (SNR), segmental signal-to-noise ratio (Seg.SNR), perceptual evaluation of speech quality (PESQ) tests and speech spectrograms are used. It is well known that the segmental SNR is more accurate in indicating the speech distortion than the overall SNR. The higher value of the segmental SNR indicates the weaker speech distortions [12]. The PESQ measures prove to be highly correlated with the subjective listening tests. The higher PESQ score indicates better perceived quality [22].

The output SNR, output Seg.SNR and PESQ improvement score of the proposed algorithm (PM) in comparison to MBSS and BSS for real-world noises and white Gaussian noises are shown in **Table 2**. From the results given in **Table 2**, we can conclude that the SNR

Table 2. Output SNR (Global), Output Seg. SNR and Perceptual evaluation of speech quality (PESQ) measure results of enhanced speech signals at (0, 5, 10, 15) dB SNRs. English sentence "*The sky that morning was clear and bright blue*" produced by a male speaker is used as original signal.

Noise Type	Enhancement Algorithms	SNR (dB)				Seg.SNR (dB)				PESQ Improvement Score			
		0 dB	5 dB	10 dB	15 dB	0 dB	5 dB	10 dB	15 dB	0 dB	5 dB	10 dB	15 dB
Car	BSS	0.95	1.40	1.59	1.71	0.86	1.17	1.28	1.33	1.749	1.925	2.154	2.213
	MBSS	2.75	7.05	9.61	12.85	2.68	6.86	9.39	12.59	1.496	1.982	2.259	2.602
	PM	**4.34**	**9.31**	**12.30**	**15.63**	**4.23**	**9.05**	**12.05**	**15.41**	**1.435**	**1.625**	**2.432**	**2.640**
Train	BSS	1.24	1.43	1.59	1.67	1.12	1.20	1.29	1.32	1.873	1.666	2.079	2.156
	MBSS	4.07	5.73	9.55	11.78	3.88	5.53	9.35	11.59	1.513	1.696	2.129	2.382
	PM	**5.32**	**7.83**	**11.80**	**14.79**	**5.12**	**7.60**	**11.59**	**14.58**	**1.522**	**1.611**	**2.143**	**2.327**
Restaurant	BSS	2.07	1.59	1.84	1.64	1.75	1.27	1.37	1.30	1.682	1.843	2.002	2.165
	MBSS	2.66	6.02	9.54	11.18	2.52	5.84	9.29	10.90	1.842	2.062	2.367	2.603
	PM	**2.35**	**7.45**	**11.52**	**14.36**	**2.14**	**7.25**	**9.96**	**14.38**	**1.633**	**1.958**	**2.321**	**2.563**
Babble	BSS	1.47	1.52	1.63	1.74	1.19	1.21	1.29	1.35	1.481	1.924	2.110	2.215
	MBSS	2.66	5.95	9.54	11.81	2.52	5.74	9.31	11.52	1.812	2.208	2.394	2.650
	PM	**3.06**	**7.45**	**11.52**	**14.36**	**2.87**	**7.17**	**11.31**	**14.11**	**1.334**	**2.105**	**2.436**	**2.657**
Airport	BSS	1.66	1.52	1.62	1.72	1.43	1.23	1.30	1.33	1.407	1.939	2.092	2.204
	MBSS	3.93	6.75	9.06	12.04	3.78	6.53	8.81	11.77	1.790	2.106	2.323	2.681
	PM	**5.31**	**7.58**	**11.62**	**15.01**	**5.26**	**7.43**	**11.44**	**14.81**	**1.462**	**2.117**	**2.424**	**2.729**
Street	BSS	2.49	1.51	1.66	1.55	1.71	1.25	1.31	1.29	1.511	1.833	2.045	2.018
	MBSS	1.88	5.60	9.42	9.93	1.71	5.39	9.22	9.73	1.592	1.933	2.249	2.213
	PM	**9.81**	**7.19**	**11.50**	**12.21**	**9.81**	**7.02**	**11.36**	**12.0**	**1.402**	**2.017**	**2.304**	**2.297**
Exhibition	BSS	2.08	1.63	1.66	1.74	2.01	1.34	1.35	1.37	1.721	1.655	2.109	2.127
	MBSS	2.24	7.18	9.09	12.36	2.05	6.99	8.92	12.13	1.527	1.977	1.968	2.517
	PM	**7.04**	**9.13**	**11.07**	**15.40**	**7.01**	**8.99**	**10.94**	**15.20**	**1.488**	**1.979**	**2.097**	**2.716**
White	BSS	1.42	1.59	1.70	1.78	1.18	1.31	1.34	1.38	1.663	1.957	2.087	2.151
	MBSS	6.10	8.80	11.93	13.46	5.90	8.63	11.77	13.26	1.655	1.971	2.303	2.563
	PM	**4.28**	**10.26**	**13.61**	**16.81**	**4.07**	**10.0**	**13.40**	**16.63**	**1.535**	**1.825**	**2.229**	**2.676**

and Seg. SNR results of the proposed algorithm is good for non-stationary and stationary noises and the PESQ improvement score is good at SNR more than 5 dB for non-stationary and stationary noises.

The objective measures do not give indications about the structure of the remnant musical noise. Speech spectrograms constitute a well-suited tool for observing this structure. **Figures 5-10** shows the speech spectrograms and temporal waveforms obtained with the PM with the value of PESQ. It can be seen from **Figures 5-10**, the musical structure of the remnant noise is reduced more by PM, even compared to MBSS algorithms. Thus,

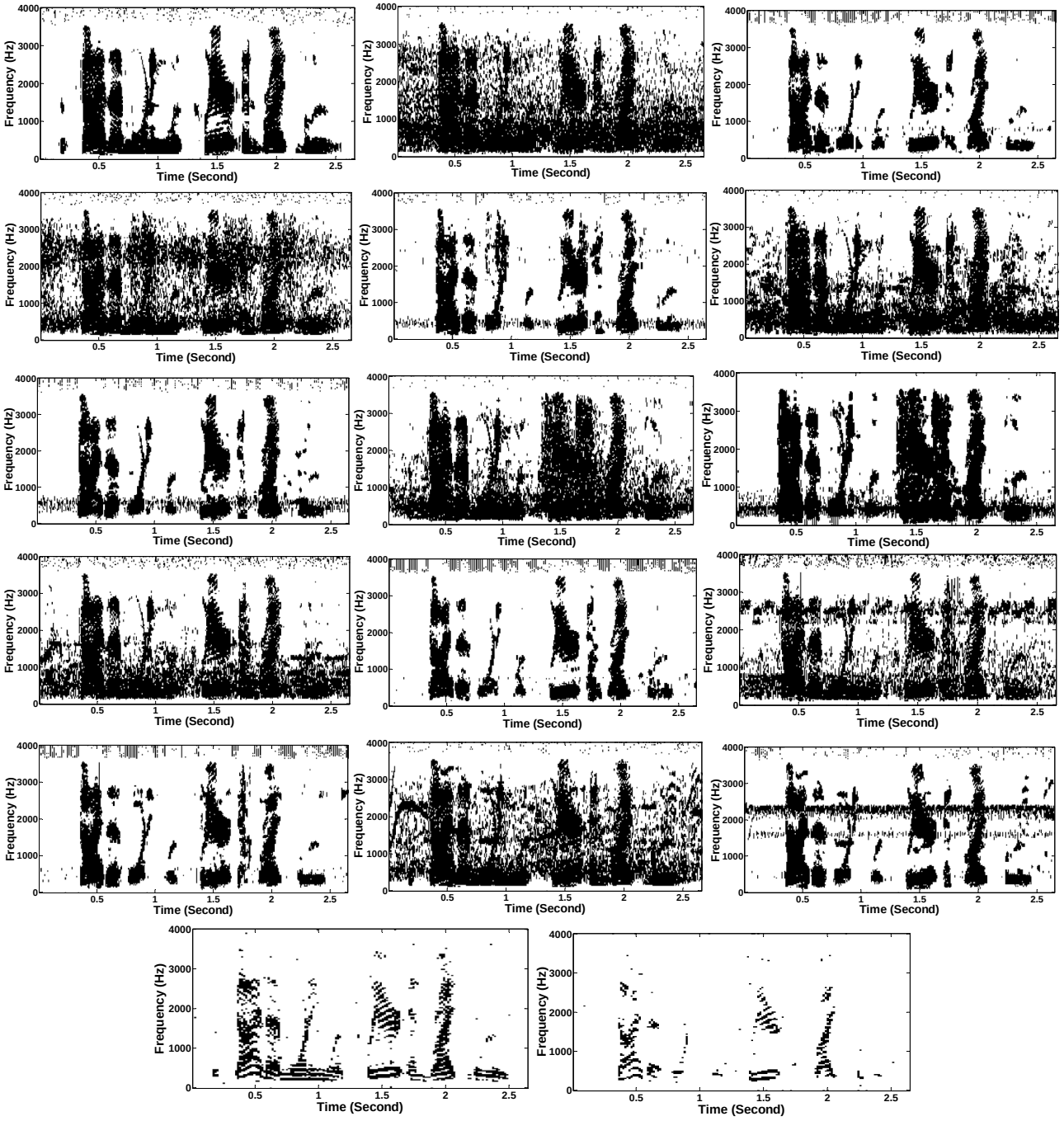


Figure 5. Speech spectrograms (From top to bottom). (a) Clean speech: sp10 utterance, "*The sky that morning was clear and bright blue*", by a male speaker from the NOIZEUS corpus; (b, d, f, h, j, l, n, p) speech degraded by car noise, train noise, babble noise, restaurant noise, airport noise, street noise, exhibition noise, and white noise respectively (10 dB SNR); and (c, e, g, i, k, m, o, q) corresponding enhanced speech.

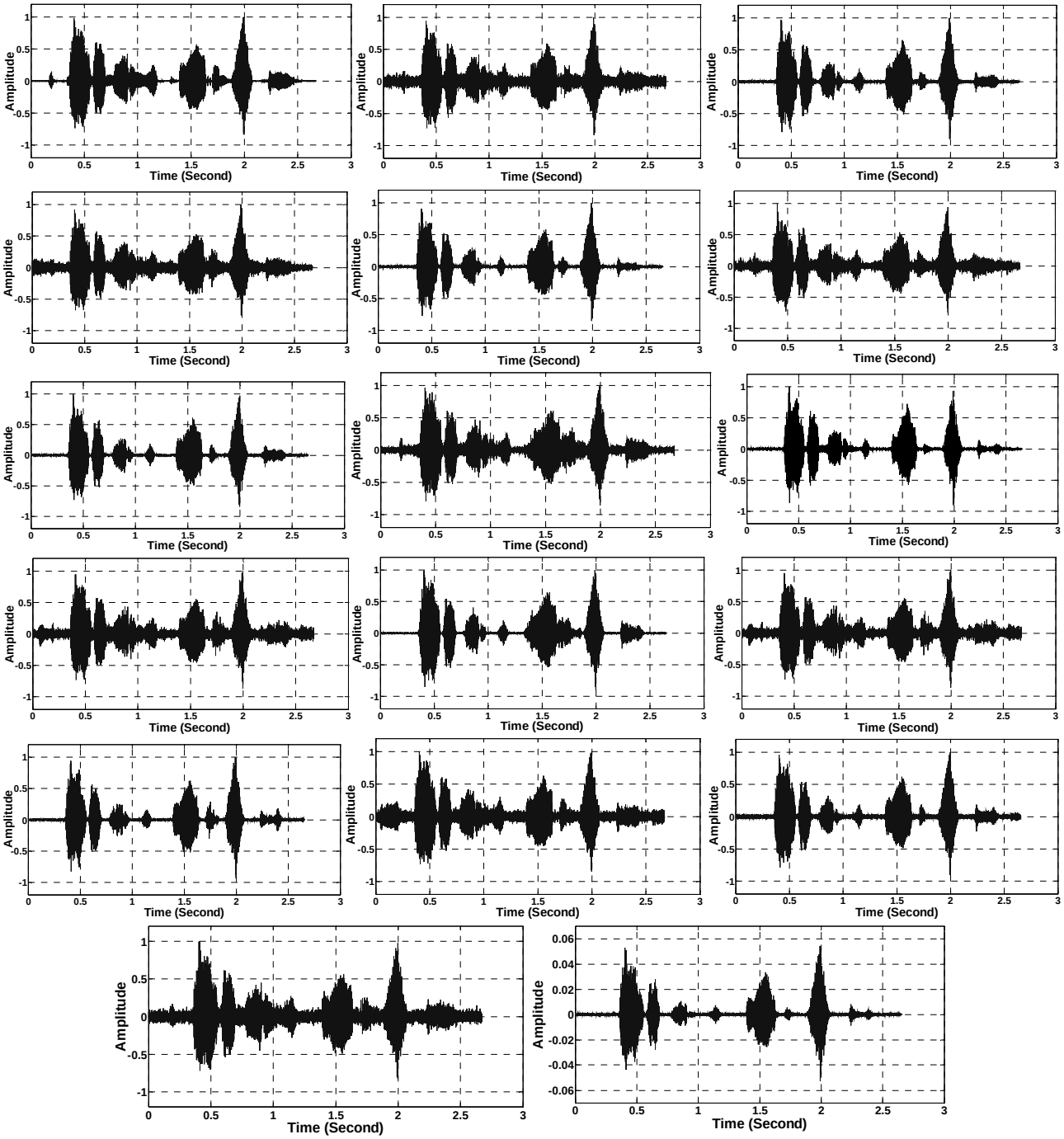


Figure 6. Temporal waveforms (From top to bottom). (a) Clean speech: sp10 utterances, "*The sky that morning was clear and bright blue*", by a male speaker from the NOIZEUS corpus; (b, d, f, h, j, l, n, p) speech degraded by car noise, train noise, babble noise, restaurant noise, airport noise, street noise, exhibition noise, and white noise respectively (10 dB SNR); and (c, e, g, i, k, m, o, q) corresponding enhanced speech.

speech enhanced with the PM is more pleasant and the remnant noise has a "perceptually white quality" while distortion remains acceptable. This confirms the values of the SNR, Seg.SNR, and PESQ; also it is validated by speech spectrogram.

5. Conclusions

In this paper, critical-band rate scale based on improved multi-band spectral subtraction algorithm is presented for enhancement of speech degraded by non-stationary noises. In the proposed enhancement algorithm, the conjuga-

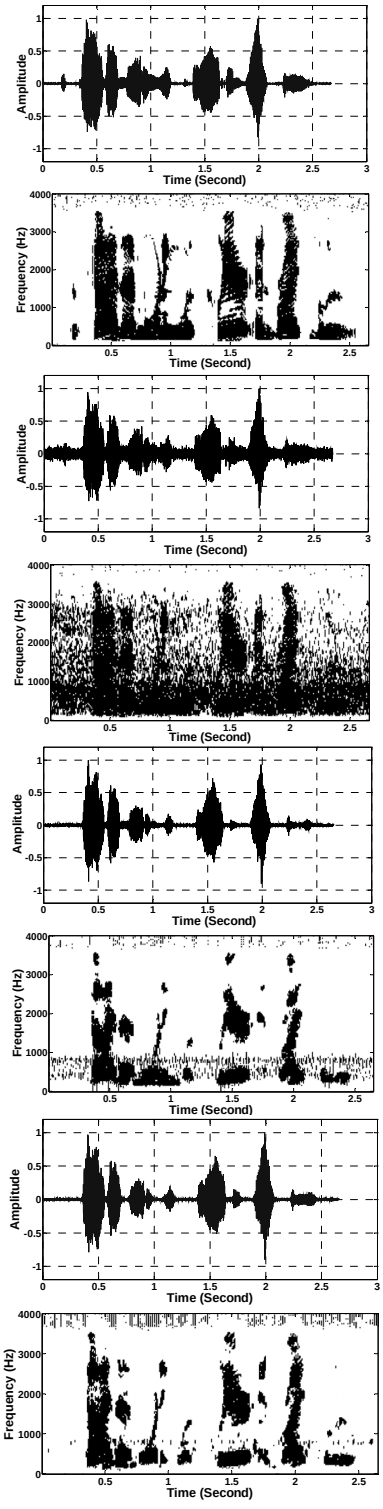


Figure 7. Temporal waveforms and speech spectrogram (From top to bottom). (a) Clean speech: sp10 utterance, "*The sky that morning was clear and bright blue*", by a male speaker from the NOIZEUS corpus; (b) speech degraded by car noise (10 dB SNR); (c) speech enhanced by MBSS (PESQ = 2.259); and (d) speech enhanced by PM (PESQ = 2.432).

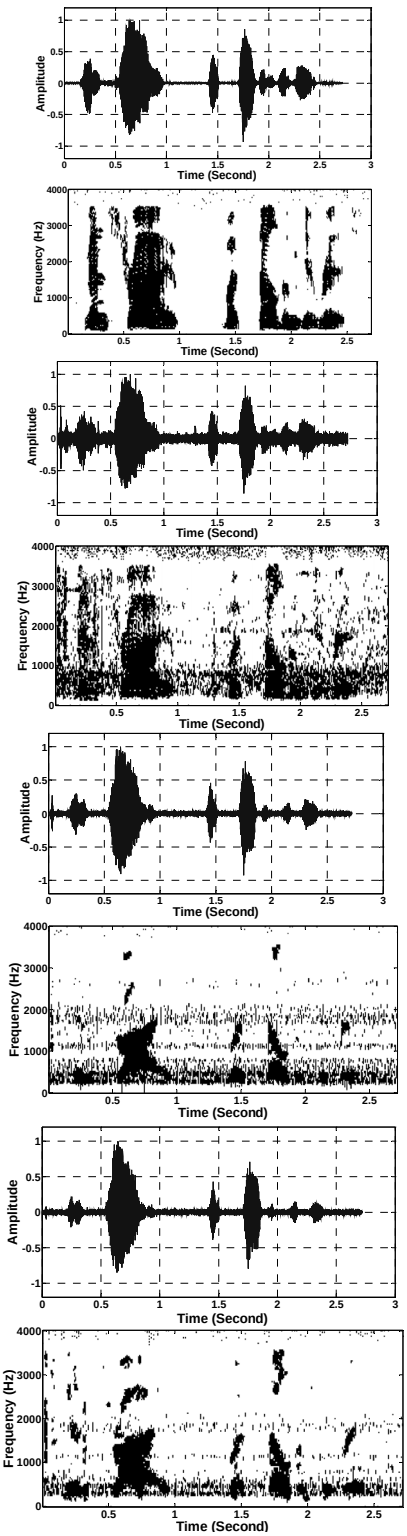


Figure 8. Temporal waveforms and speech spectrogram (From top to bottom). (a) Clean speech: sp6 utterance, "*Men strive but seldom get rich*", by a male speaker from the NOIZEUS corpus; (b) speech degraded by car noise (10 dB SNR); (c) speech enhanced by MBSS (PESQ = 2.157); and (d) speech enhanced by PM (PESQ = 2.330).

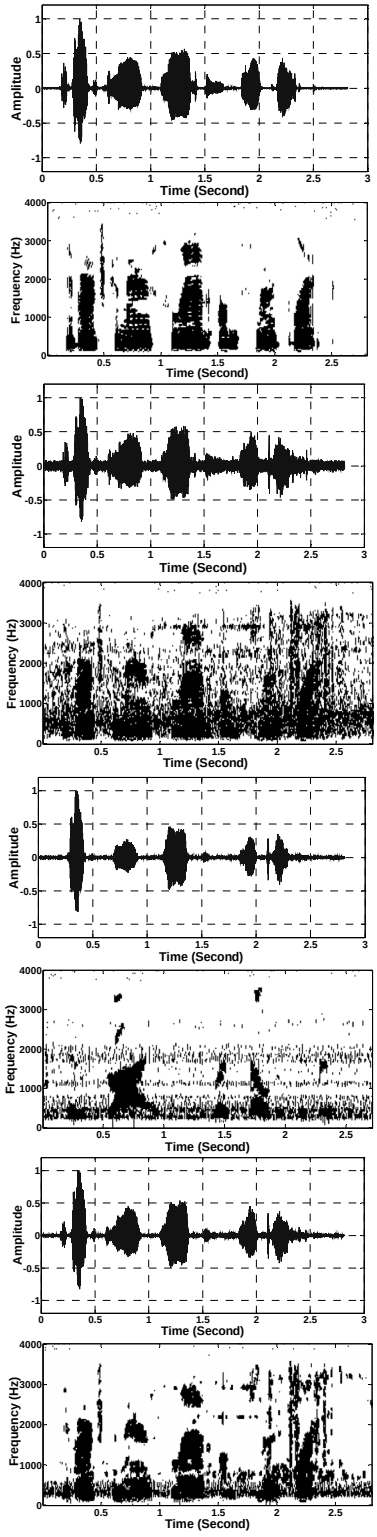


Figure 9. Temporal waveforms and speech spectrogram (From top to bottom). (a) clean speech: sp1 utterance, "*The birch canoe slid on the smooth planks*", by a male speaker from the NOIZEUS corpus; (b) speech degraded by car noise (10 dB SNR); (c) speech enhanced by MBSS (PESQ = 2.030); and (d) speech enhanced by PM (PESQ = 2.167).

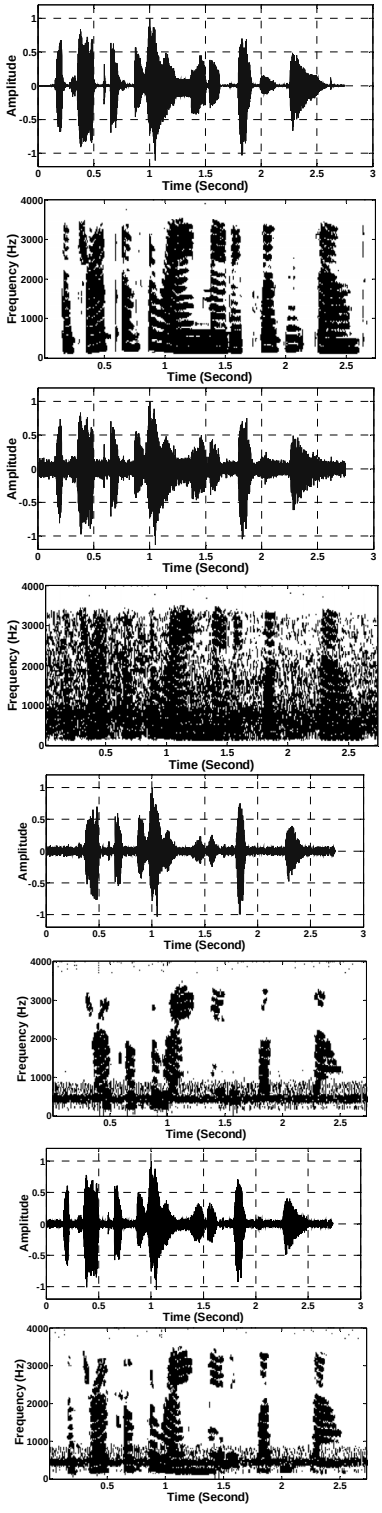


Figure 10. Temporal waveforms and speech spectrogram (From top to bottom). (a) clean speech: sp12 utterance, "*The drip of the rain made a pleasant sound*", by a female speaker from the NOIZEUS corpus; (b) speech degraded by car noise (10 dB SNR); (c) speech enhanced by MBSS (PESQ = 2.005); and (d) speech enhanced by PM (PESQ = 2.483).

tive three-three critical-bands are grouped together into six non-uniformly spaced frequency bands that are closely approximating with the non-uniform frequency spacing given by the human auditory system. Additionally, the noise is estimated from each non-uniform spaced frequency band by an adaptive noise estimation approach that does not need speech pause detection.

The simulation results and evaluations tests with the number of non-stationary and a stationary noises reveal that the proposed algorithm suppresses the remnant noise tones efficiently that appear in case of the basic spectral sub- traction and standard multi-band spectral subtraction algorithm, also improves the overall quality of degraded speech at low SNRs. Moreover, the proposed algorithm has strong flexibility to adapt any complicated rigorous speech environment by adjusting the over-subtraction factor for each non-uniformly spaced frequency band separately.

REFERENCES

[1] D. O'Shaughnessy, "Speech Communications: Human and Machine," 2nd Edition, University Press (India) Pvt. Ltd., Hyderabad, 2007.

[2] Y. Ephraim, "Statistical-Model-Based Speech Enhancement Systems," *Proceedings of the IEEE*, Vol. 80, No. 10, 1992, pp. 1526-1555. doi:10.1109/5.168664

[3] Y. Ephraim, H. L. Ari and W. Roberts, "A Brief Survey of Speech Enhancement," In: *The Electrical Engineering Handbook*, 3rd Edition, CRC, Boca Raton, 2006.

[4] Y. Ephraim and I. Cohen, "Recent Advancements in Speech Enhancement," In: *The Electrical Engineering Handbook*, CRC Press, Boca Raton, 2006, pp. 12-26.

[5] J. S. Lim and A. V. Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech," *Proceedings of the IEEE*, Vol. 67, No. 12, 1979, pp. 1586-1604. doi:10.1109/PROC.1979.11540

[6] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Transactions on Acoustic, Speech, and Signal Processing*, Vol. 27, No. 2, 1979, pp. 113-120. doi:10.1109/TASSP.1979.1163209

[7] M. Berouti, R. Schwartz and J. Makhoul, "Enhancement of Speech Corrupted by Acoustic Noise," *Proceedings of International Conference on Acoustic, Speech, and Signal Processing*, Washington DC, April 1979, pp. 208-211.

[8] S. Kamath and P. Loizou, "A Multi-Band Spectral Subtraction Method for Enhancing Speech Corrupted by Colored Noise," *Proceedings of International Conference on Acoustic, Speech, and Signal Processing*, Orlando, 13-17 May 2002. doi:10.1109/ICASSP.2002.5745591

[9] R. M. Udrea, N. Vizireanu, S. Ciochina and S. Halunga, "Non-Linear Spectral Subtraction Method for Colored Noise Reduction Using Multi-Band Bark Scale," *Signal Processing*, Vol. 88, No. 5, 2008, pp. 1299-1303. doi:10.1016/j.sigpro.2007.11.023

[10] S. Li, J. Q. Wang and X. J. Jing, "The Application of Non-Linear Spectral Subtraction Method on Millimeter Wave Conducted Speech Enhancement," *Mathematical Problems in Engineering*, Vol. 2010, 2010, Article ID: 570940. doi:10.1155/2010/570940

[11] V. Rama Rao, R. Murthy and K. S. Rao, "Speech Enhancement Using Cross-Correlation Compensated Multi-Band Wiener Filter Combined with Harmonic Regeneration," *Journal of Signal and Information Processing*, Vol. 2, No. 2, 2011, pp. 117-124. doi:10.4236/jsip.2011.22016

[12] P. C. Loizou, "Speech Enhancement: Theory and Practice," Taylor and Francis, 2007.

[13] R. Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," *IEEE Transactions on Speech and Audio Processing*, Vol. 9, No. 5, 2001, pp. 504-512. doi:10.1109/89.928915

[14] I. Cohen, "Noise Spectrum Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging," *IEEE Transactions on Speech and Audio Processing*, Vol. 11, No. 5, 2003, pp. 466-475. doi:10.1109/TSA.2003.811544

[15] G. Doblinger, "Computationally Efficient Speech Enhancement by Spectral Minima Tracking in Subbands," *Proceedings of Euro Speech*, Vol. 2, 1995, pp. 1513-1516.

[16] L. Lin, W. H. Holmes and E. Ambikairajah, "Adaptive Noise Estimation Algorithm for Speech Enhancement," *Electronics Letters*, Vol. 39, No. 9, 2003, pp. 754-755. doi:10.1049/el:20030480

[17] E. Zwicker and H. Fastl, "Psychoacoustics: Facts and Models," Springer-Verlag, Berlin, 1990.

[18] E. Zwicker and E. Terhardt, "Analytical Expressions for Critical-Band Rate and Critical Bandwidth as a Function of Frequency," *Journal of Acoustic Society of America*, Vol. 68, No. 5, 1980, pp. 1523-1525. doi:10.1121/1.385079

[19] N. Upadhyay and A. Karmakar, "A Perceptually Motivated Multi-Band Spectral Subtraction Algorithm for Enhancement of Degraded Speech," *Proceedings of IEEE International Conference on Computer, and Communication Technology*, Allahabad, 23-25 November 2012, pp. 340-345.

[20] N. Upadhyay and A. Karmakar, "An Auditory Perception Based Improved Multi-Band Spectral Subtraction Algorithm for Enhancement of Speech Degraded by Non-Stationary Noises," *Proceedings of IEEE International Conference on Intelligent Human Computer Interaction*, Kharagpur, 27-29 December 2012, pp. 392-398.

[21] "A Noisy Speech Corpus for Assessment of Speech Enhancement Algorithms." http://www.utdallas.edu/~loizou/speech/noizeus/

[22] "Perceptual Evaluation of Speech Quality (PESQ), and Objective Method for End-to-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs," ITU, ITU-T Rec. P. 862, 2000.

