Journal of Signal and Information Processing, **2013, 4, 197-211**
http://dx.doi.org/10.4236/jsip.2013.42027 Published Online May 2013 (http://www.scirp.org/journal/jsip)

A Multi-Band Speech Enhancement Algorithm Exploiting Iterative Processing for Enhancement of Single Channel Speech

Navneet Upadhyay[1], Abhijit Karmakar[2]

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

Received January 20th, 2013; revised February 22nd, 2013; accepted March 3rd, 2013

ABSTRACT

This paper proposes a multi-band speech enhancement algorithm exploiting iterative processing for enhancement of single channel speech. In the proposed algorithm, the output of the multi-band spectral subtraction (MBSS) algorithm is used as the input signal again for next iteration process. As after the first MBSS processing step, the additive noise transforms to the remnant noise, the remnant noise needs to be further re-estimated. The proposed algorithm reduces the remnant musical noise further by iterating the enhanced output signal to the input again and performing the operation repeatedly. The newly estimated remnant noise is further used to process the next MBSS step. This procedure is iterated a small number of times. The proposed algorithm estimates noise in each iteration and spectral over-subtraction is executed independently in each band. The experiments are conducted for various types of noises. The performance of the proposed enhancement algorithm is evaluated for various types of noises at different level of SNRs using, 1) objective quality measures: signal-to-noise ratio (SNR), segmental SNR, perceptual evaluation of speech quality (PESQ); and 2) subjective quality measure: mean opinion score (MOS). The results of proposed enhancement algorithm are compared with the popular MBSS algorithm. Experimental results as well as the objective and subjective quality measurement test results confirm that the enhanced speech obtained from the proposed algorithm is more pleasant to listeners than speech enhanced by classical MBSS algorithm.

Keywords: Speech Enhancement; Multi-Band Spectral Subtraction; Iterative Processing; Remnant Musical Noise

1. Introduction

Speech is the most prominent and primary mode of interaction between human-to-human and human-to-machine communications in various fields such as automatic speech recognition and speaker identification [1]. The present day speech communication systems are severely degraded due to various types of interfering signals which make the listening task difficult for a direct listener and cause inaccurate transfer of information [2]. Therefore, to obtain near-transparent speech communication in applications such as in mobile phones, noise suppression or enhancement of degraded speech is one of the main research endeavors in the field of speech signal processing over the last few decades. The main focus of speech enhancement research is to minimize the degree of distortion of the desired speech signal and to improve one or more perceptual aspects of speech, such as the speech quality and/or intelligibility of the processed speech [3,4]. These two features, quality and intelligibility, are however, uncorrelated and independent of 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 high quality speech may be low in intelligibility while a low quality speech may be high in intelligibility [5].

The classification of speech enhancement methods depend on the number of microphones that are used for collecting speech such as single, dual or multi-channel. Although the performance of multi-channel speech enhancement is better than single channel speech enhancement [1], the single channel speech enhancement is still a

significant field of research interest because of its simple implementation and easy computation. Single channel speech enhancement method uses only one microphone to collect noisy data but no additional information about the degrading noise and the clean speech is available. The estimation of the spectral magnitude from the noisy speech is easier than the estimate of both magnitude and phase. In [6], 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 proposed by Boll [7], is one of the most widely used methods based on the direct estimation of STSM. The main attraction of spectral subtraction method is: 1) Its relative simplicity, in that it only requires an estimate of the noise spectrum, and 2) Its high flexibility against the variation of subtraction parameters. Despite its capability of removing background noise, spectral subtraction [7] introduces perceptually noticeable spectral artifacts, known as remnant musical noise, which is composed of un-natural artifacts with random frequencies and perceptually annoys the human ear. This noise is caused due to the inaccuracies in the short-time noise spectrum estimate and it faces difficulties in pause detection. In recent years, a number of speech enhancement algorithms have been developed to deal with the modifications of the spectral subtraction method to combat the problem of remnant musical noise artifacts and improve the quality of speech in noisy environments. In [7], magnitude averaging rule is proposed. In [8], the over-subtraction of noise is proposed and defined a spectral floor to make remnant musical noise inaudible. In [9], a speech enhancement algorithm by incorporating the multi-band model in frequency domain is proposed.

This paper proposes a novel algorithm for suppressing the remnant noise and enhancement of single channel speech. In the proposed algorithm, the output of multi-band spectral subtraction (MBSS) is used as the input signal again for next iteration process. After the MBSS algorithm, the additive noise is transformed to remnant noise. The remnant noise is re-estimated in each iteration and spectral over-subtraction executed separately in each band. This procedure is iterated a small number of times. The performance of enhanced speech is characterized by a trade-off between the amount of noise reduction, speech distortion, and the level of remnant noise.

The rest of paper is structured as follows. Section 2 describes the principle of spectral subtraction method for speech enhancement [7], the spectral over-subtraction (SOS) [8], and MBSS [9] which serve as a reference of our proposed algorithm platform. In Section 3, the proposed enhancement algorithm, multi-band spectral subtraction algorithm exploiting iterative processing (IP-MBSS) is introduced for suppression of remnant musical noise. Section 4 reports the experimental results and performance evaluation. The conclusion is drawn in Section 5.

2. Principle of Spectral Subtraction Method

The spectral subtraction is one of the most popular and computationally simple methods for effectively suppressing the background noise from the noisy speech as it involves a single forward and inverse transform. The first comprehensive spectral subtraction method, proposed by Boll [7], is based on non-parametric approach, which simply needs an estimate of noise spectrum and used for both speech enhancement and recognition.

In real-world listening environments, the speech signal is mostly corrupted by additive noise [3,7]. Additive noise is typically the background noise and is uncorrelated with the clean speech signal. The background noise can be of stationary type, such as white Gaussian noise (WGN) or of non-stationary or colored type. The speech degraded by background noise is termed as noisy speech. The noisy speech can be modeled as the sum of the clean speech and the random noise [3,7] 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 or quasi-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 [3,7],

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

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 termed 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 auditory system (HAS) [6]. Therefore, the STSM of noisy speech is equal to the sum of STSM of clean speech and STSM of random noise without the information of phase and (2) can be expressed as

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

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

$$|Y_w(\omega)|^2 = |S_w(\omega)|^2 + |D_w(\omega)|^2 + S_w(\omega)D_w^*(\omega) + S_w^*(\omega)D_w(\omega) \quad (4)$$

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

$$|\hat{S}_w(\omega)|^2 = |Y_w(\omega)|^2 - |\hat{D}_w(\omega)|^2 \quad (5)$$

where $|\hat{S}_w(\omega)|^2$ and $|Y_w(\omega)|^2$ is the power spectrum of estimated speech and the noisy speech, respectively. The $|\hat{D}_w(\omega)|^2$ is the average noise power, normally estimated during speech pauses.

In spectral subtraction method, it is assumed that the speech signal is degraded by additive white Gaussian noise (AWGN) with flat spectrum. In this method, the subtraction process needs to be carried-out carefully 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 HWR or full-wave rectification (FWR) is introduced. Thus, the complete power spectral subtraction algorithm is given by

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

As the human perception is insensitive to phase [6], the enhanced speech spectrum can be obtained with phase of noisy speech and the enhanced speech is reconstructed by taking the inverse STFT (ISTFT) of the enhanced spectrum using the phase of the noisy speech and overlap-add (OLA) method, can be expressed as

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

On the contrary, a generalized form of spectral subtraction method (5) can be obtained by altering the power exponent from 2 to b, which determines the sharpness of the transition.

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

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. When the noise estimate is less than perfect, two major problems occur, remnant residual noise, referred as musical noise, and speech distortion.

2.1. Spectral Over-Subtraction Algorithm

An improved version of spectral subtraction method was proposed in [8] to minimize the annoying musical noise and speech distortion. In this algorithm, the spectral subtraction method [7] uses two additional parameters, namely, over-subtraction factor, and noise spectral floor parameter [8]. The algorithm is given as

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

with $\alpha \geq 1$ and $0 \leq \beta \ll 1$.

The over-subtraction factor α controls the amount of noise power spectrum subtracted from the noisy speech power spectrum in each frame and spectral floor parameter β prevent the resultant spectrum from going below a preset minimum level rather than setting to zero. The over-subtraction factor depends on *a-posteriori* 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 $\alpha_0\left(\alpha_0 \approx 4\right)$ 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 α and β give rise to a trade-off between the amount of remnant noise and the level of perceived musical noise. For large value of 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 noise, but background noise remains present and enhanced speech is distorted.

2.2. Multi-Band Spectral Subtraction Algorithm

In real-world listening environment, the noise does not affect the speech signal uniformly over the whole spectrum. Here, some frequencies are affected more adversely than others, which eventually mean that this kind of noise is non-stationary or colored.

To take into account the fact that real-world noise affects the speech spectrum differently at various frequencies, a multi-band linear frequency spacing approach to spectral over-subtraction was presented in [9], which is the non-linear spectral subtraction approach.

In this scheme, the noisy speech spectrum is divided into $K\left(K = 4\right)$ non-overlapping uniformly spaced frequency bands, and spectral over-subtraction is applied independently in each band. The multi-band spectral subtraction algorithm re-adjusts the over-subtraction factor in each band. Thus, the estimate of the clean speech spectrum in the i^{th} band is obtained by

$$\left|\hat{S}_i\left(\omega\right)\right|^2 = \begin{cases} \left|Y_i\left(\omega\right)\right|^2 - \alpha_i \cdot \delta_i \cdot \left|\hat{D}_i\left(\omega\right)\right|^2, \text{if } \left|\hat{S}_i\left(\omega\right)\right|^2 > \beta \cdot \left|Y_i\left(\omega\right)\right|^2 \\ \beta \cdot \left|Y_i\left(\omega\right)\right|^2, \qquad \text{else} \end{cases}, \quad k_i < \omega < k_{i+1} \tag{11}$$

where k_i and k_{i+1} are the beginning and ending frequency bins of the i^{th} frequency band, α_i is the band specific over-subtraction factor of the i^{th} band, which is the function of segmental SNR of the i^{th} frequency band $\left(\text{SNR}_i\right)$ and provides a degree of control over the noise subtraction level in each band. The segmental SNR of the i^{th} frequency band $\left(\text{SNR}_i\right)$ can be calculated as

$$\text{SNR}_i\left(\text{dB}\right) = 10\log_{10}\left(\frac{\sum_{\omega=k_i}^{k_{i+1}}\left|Y_i\left(\omega\right)\right|^2}{\sum_{\omega=k_i}^{k_{i+1}}\left|\hat{D}_i\left(\omega\right)\right|^2}\right) \tag{12}$$

The band specific over-subtraction can be calculated using **Figure 1** and the value of parameters is given in Section 2.1 as

$$\alpha_i = \begin{cases} \alpha_{\max}, \\ \text{if } \text{SNR}_i \le \text{SNR}_{\min} \\ \alpha_{\max} + \left(\text{SNR}_i - \text{SNR}_{\min}\right)\left(\frac{\alpha_{\min} - \alpha_{\max}}{\text{SNR}_{\max} - \text{SNR}_{\min}}\right), \\ \text{if } \text{SNR}_{\min} \le \text{SNR}_i \le \text{SNR}_{\max} \\ \alpha_{\min}, \\ \text{if } \text{SNR}_i \ge \text{SNR}_{\max} \end{cases} \tag{13}$$

The result of an implementation of four band MBSS [9] with estimated segmental SNR of four frequency bands {60 Hz ~ 1 kHz (Band1), 1 kHz ~ 3 kHz (Band2), 2 kHz ~ 3 kHz (Band3), 3 kHz ~ 4 kHz (Band4)} of noisy speech spectrum is shown in **Figure 2**. It can be seen

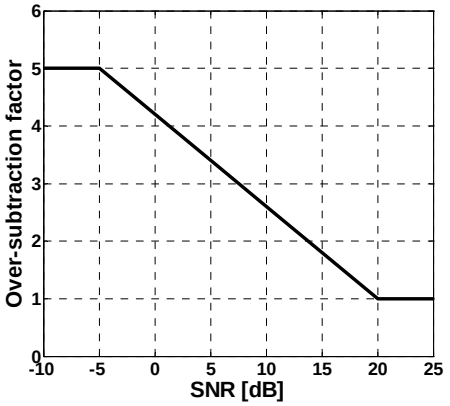


Figure 1. The relation between over-subtraction factor and SNR.

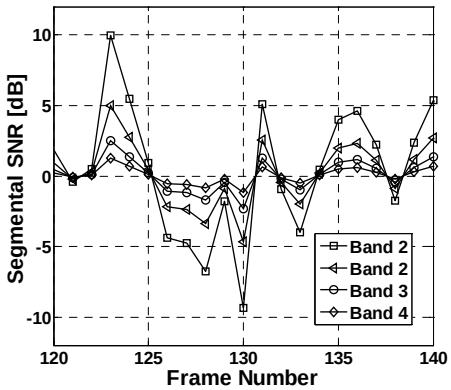


Figure 2. The segmental SNR of four linearly spaced frequency bands of degraded speech.

from the figure that the segmental SNR of the low frequency bands (Band1) is significantly higher than the segmental SNR of the high frequency bands (Band4) [9].

The δ_i is an additional band subtraction factor that can be individually set for each frequency band to customize the noise removal process and provide an additional degree of control over the noise subtraction level in each band. The values of δ_i [9] is empirically calculated as most of the speech energy is concentrated below 1 kHz and set to

$$\delta_i = \begin{cases} 1, & f_i \le 1\,\text{kHz} \\ 2.5, & 1\,\text{kHz} < f_i \le \dfrac{f_s}{2} - 2\,\text{kHz} \\ 1.5, & f_i > \dfrac{f_s}{2} - 2\,\text{kHz} \end{cases} \tag{14}$$

Here f_i is the upper bound frequency of the i^{th} band and f_s is the sampling frequency. The motivation for using smaller values of δ_i for the low frequency bands is to minimize speech distortion, since most of the speech energy is present in the lower frequencies. Both factors, α_i and δ_i can be adjusted for each band for different speech conditions to get better speech quality.

As the real-world noise is highly random in nature, improvement in the MBSS algorithm for reduction of WGN is necessary. But the performance of MBSS algorithm is better than spectral subtraction method [7] and SOS algorithm [8]. The block diagram of MBSS algorithm is shown in [10].

3. Multi-Band Spectral Subtraction Exploiting Iterative Processing

In order to reduce the remnant musical noise, produced by the multi-band spectral subtraction algorithm, we have used the MBSS algorithm [9] that makes use of the iterative processing [11]. The iterative processing is a technique in which the speech enhancement procedure is executed on the estimated speech that is taken as the input and processed repeatedly to obtain the further enhanced speech and thus reducing the remnant noise. Therefore, the reduction of remnant musical noise can be achieved by estimating noise from processed speech in each iteration and determines the quality and intelligibility of the enhanced speech. The iterative method is motivated by Wiener filtering method [6,11,12] which is one of the speech enhancement techniques.

If we regard the process of noise estimation and the MBSS as a filtering step, then the output signal of the filter is used not only for designing the filter but also as the input signal of the next iteration process. More importantly, this filter can be refreshed adaptively by re-estimating the remnant noise to improve the speech quality and intelligibility effectively [11]. The block diagram of iterative processing based multi-band spectral subtraction algorithm (IP-MBSS) is illustrated in **Figure 3**.

If m denotes the iterations number, then let us assume that the noisy speech signal at the m^{th} iteration step is given by

$$y(m,n) = s(m,n) + d(m,n), n \in (0, N-1) \tag{15}$$

Here, $y(m, n)$, $s(m, n)$, and $d(m, n)$ are the n^{th} sample at m^{th} iteration step of the discrete-time signal of noisy speech, clean speech and the noise respectively. The m^{th} iteration step of the MBSS algorithm is obtained as

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

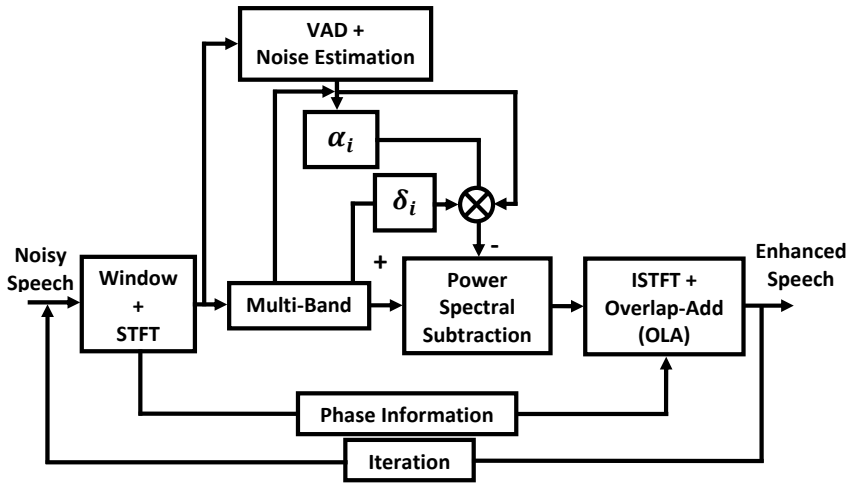


Figure 3. Block diagram of multi-band spectral subtraction exploiting iterative processing algorithm.

where $$k_i < \omega < k_{i+1} \quad (16)$$

$$\left|\hat{S}_i(m+1,\omega)\right|^2 = \left|\hat{S}_i(m,\omega)\right|^2 \cdot \left|Y_i(m,\omega)\right|^2 \quad (17)$$

where $\left|\hat{S}_i(m,\omega)\right|^2$, $\left|Y_i(m,\omega)\right|^2$, and $\left|\hat{D}_i(m,\omega)\right|^2$ is the power spectrum of estimated speech, noisy speech and estimated noise power in the i^{th} band at the m^{th} iteration step, respectively. In the $(m+1)^{\text{th}}$ iteration processing, the output signal $\hat{S}_i(m,\omega)$ obtained after the m^{th} iteration is set as the input signal again as

$$y(m+1,n) = \hat{s}(m,n) \quad (18)$$

In this algorithm, the noise spectrum, that is used for each iteration, is estimated from the noise component that remained after the iterative processing of the previous stage. Here, the noise component of $y(m+1,n)$ becomes the remnant noise component that could not be suppressed by the MBSS at m^{th} iteration. As the amount of the noise component is reduced in each MBSS processing step, increasing the number of iterations in this method will reduce the amount of noise, progressively.

The number of iteration steps is the most important parameter of this algorithm which affects the performance of the speech enhancement system [11,13]. The segmental SNR at the end of each iteration step depends on over-subtraction factor α and it increases with the number of iterations.

4. Experimental Results and Performance Evaluation

This section presents the experiments results and performance evaluation of the proposed enhancement algorithm as well as a comparison with the conventional MBSS algorithm. For simulations, we have employed MATLAB software as the simulation environment. The clean speech and noisy speech samples have been taken from NOIZEUS corpus speech database [14]. The NOIZEUS database is composed of 30 phonetically-balanced sentences belonging to six speakers, three male and three female, degraded by seven different real-world noises at different levels of SNRs. A total of four different utterances pronounced by male speakers and female speaker are used in our evaluation.

Noise signals have different time-frequency distributions, and therefore a different impact on speech. For our purpose, the sentences are degraded with seven types real-world noises and white Gaussian noise, at varying SNR levels *i.e.* 0 dB to 15 dB in steps of 5 dB. The real-world noises are car, train, restaurant, babble, airport, street, and exhibition. The performance of the proposed enhancement algorithm is tested on such noisy speech samples.

For our enhancement experiments, the 8 kHz sampled speech signals are quantized into digital signal with 16-bit resolution. The frame size is chosen to be 256 (32 ms), with 50% overlapping. The sinusoidal Hamming window with size 256 is applied to the noisy signal. The noise estimate is updated during the silence frames by using averaging (20 frames) with the value of smoothing factor for noise power spectral density estimation is 0.9.

The iteration time is an important factor of the proposed algorithm, IP-MBSS, which effects on the performance of speech enhancement. In order to explore the relationship between the performance of speech enhancement and the iteration times, the variation of the mean over-subtraction factor (α) of the car speech with iteration times are shown in **Figure 4**. It can be seen from the figure that the α increases as the iteration number increases, which suggest the larger iteration number corresponds to better speech enhancement with less remnant noise. However, both the speech waveforms and the speech spectrogram suggest that the larger iteration number would eliminate part of the normal speech component to some extent while it works effectively for reducing the remnant noise. Therefore, the iteration number for the car speech is set to 2 to 3 and the value of other parameters have been taken as same as the reference algorithm, MBSS. The signal waveforms and spectrograms of clean, noisy and enhanced speech signals were given in **Figures 5-11**.

To evaluate the performance of proposed enhancement algorithm, the objective quality and subjective quality measures are used. The objective quality measure are SNR, segmental SNR (Seg.SNR), and perceptual evaluation of speech quality (PESQ) while the subjective measure is the mean opinion score (MOS).

4.1. Objective Measure

1) Signal-to-noise ratio: SNR is defined as the ratio of the total signal energy to the total noise energy in the utterance. The following equation is used for evaluation of SNR results of enhanced speech signals

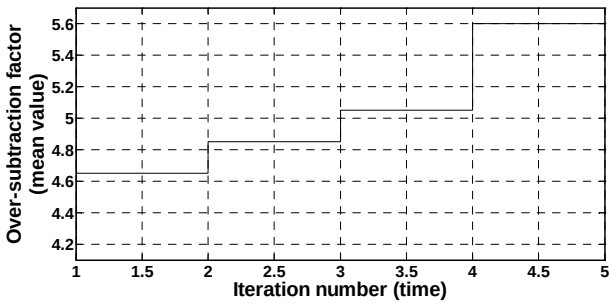
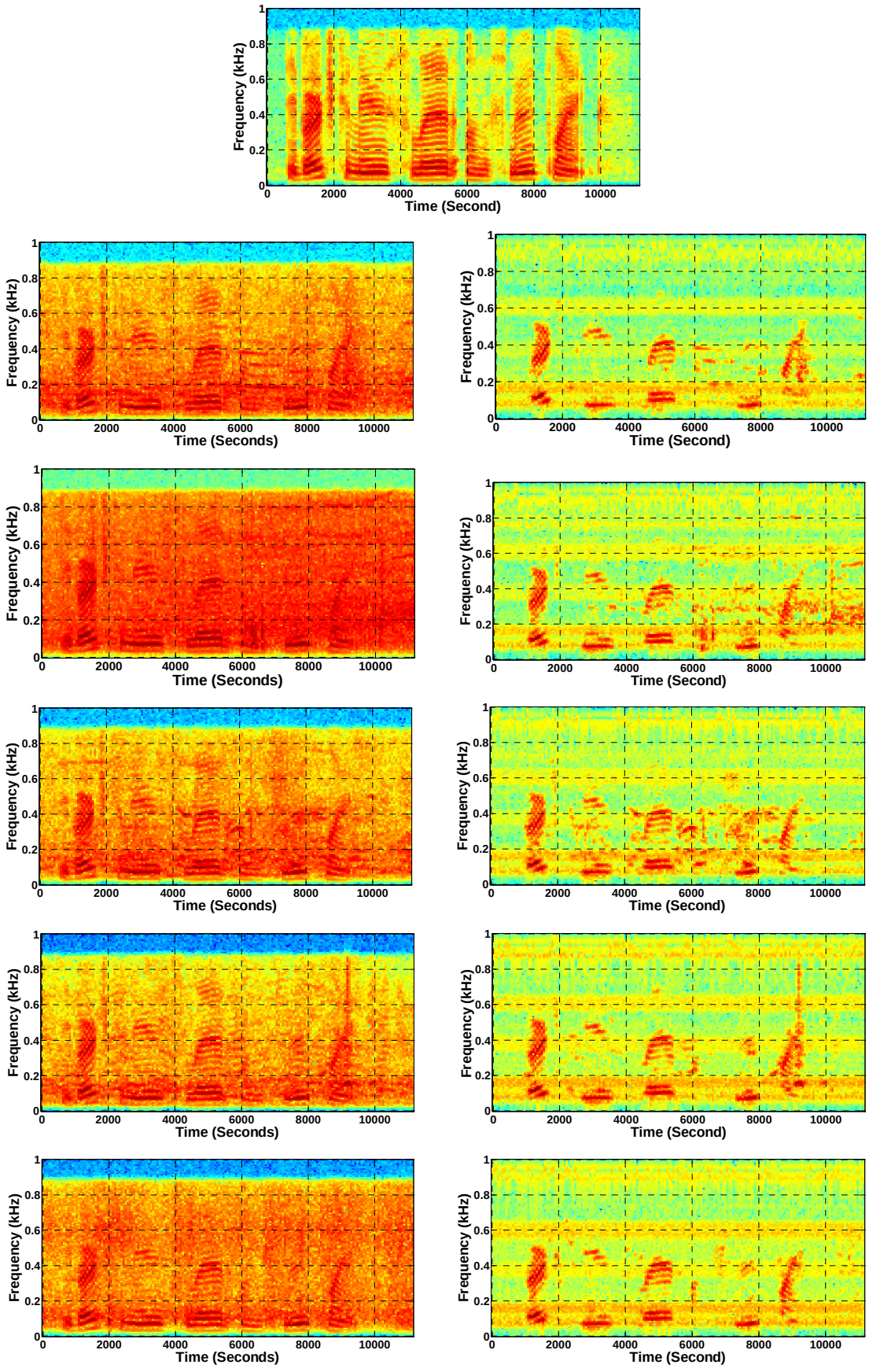


Figure 4. Variations of over-subtraction factor (mean value) with iteration times (number).



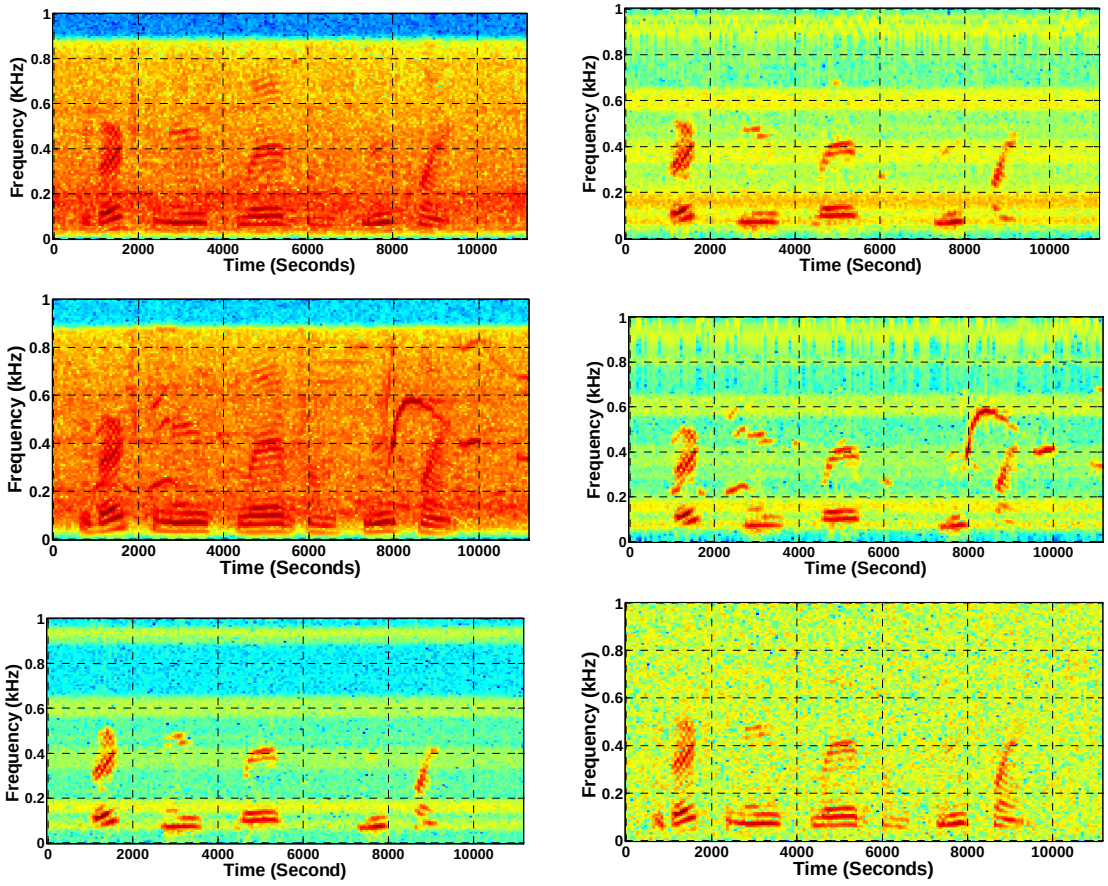


Figure 5. Speech spectrograms of sp1 utterance, "*The birch canoe slid on the smooth planks*", by a male speaker from the NOIZEUS speech corpus: (from top to bottom) clean speech; (left side, from top to bottom) speech degraded by car noise, train noise, babble noise, restaurant noise, airport, street, exhibition, and white noise, respectively (5 dB SNR); (right side, from top to bottom) corresponding enhanced speech.

$$\mathrm{SNR} = 10\log_{10}\left(\frac{\sum_{n=1}^{L} s^2(n)}{\sum_{n=1}^{L}\left\{s(n)-\hat{s}(n)\right\}^2}\right) \tag{19}$$

where $s(n)$ is the clean speech signal, $\hat{s}(n)$ is the enhanced speech reproduced by a speech processing system, n is the sample index, and L is the number of samples in both speech signals. The summation is performed over the signal length.

2) Segmental signal-to-noise ratio: Seg.SNR is the average ratio of signal energy to noise energy per frame, and can be expressed as follows:

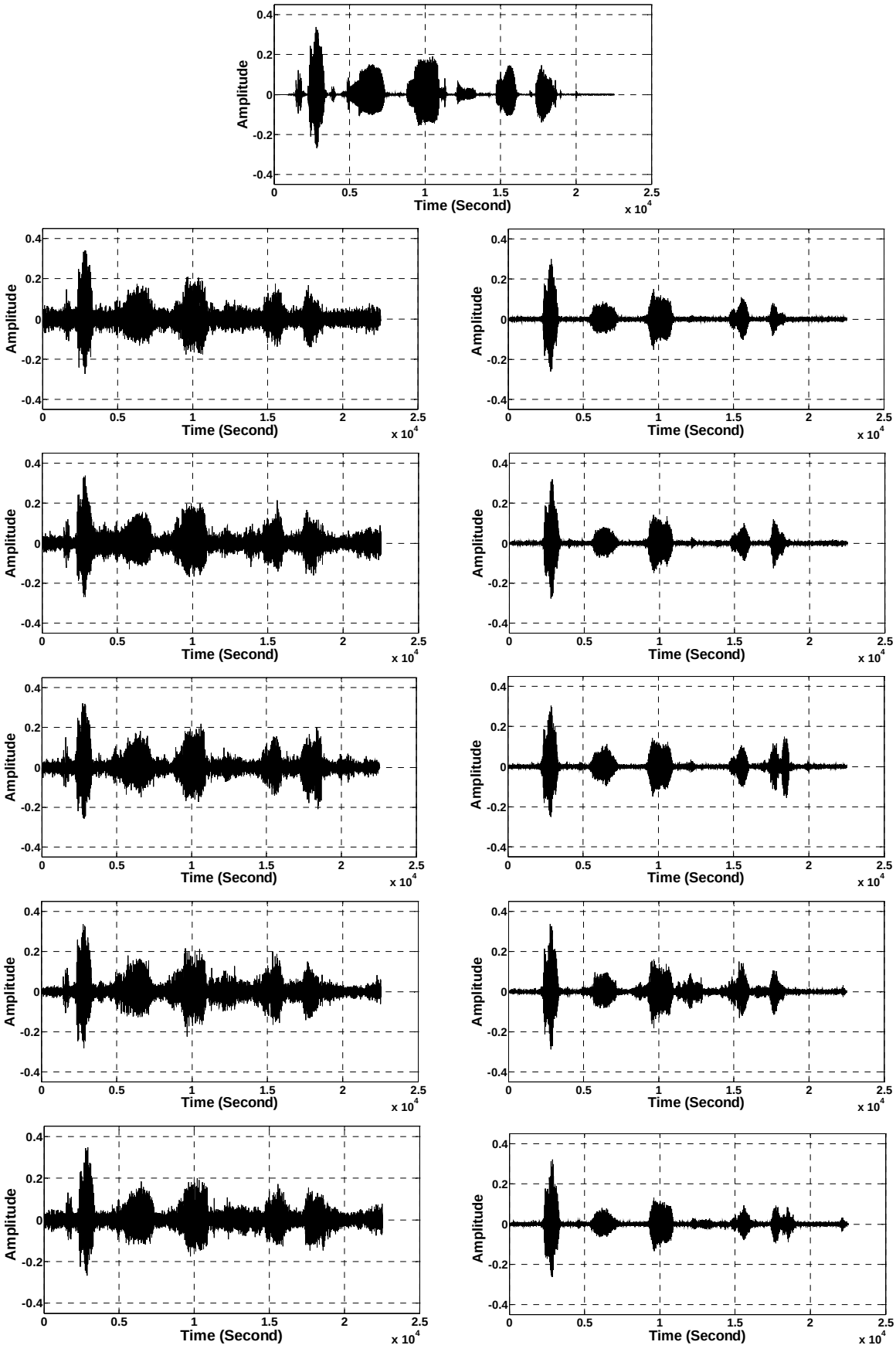
$$\mathrm{Seg.SNR} = \frac{10}{M}\sum_{m=0}^{M-1}\log_{10}\left(\frac{\sum_{n=N_m}^{N_m+N-1} s^2(n)}{\sum_{n=N_m}^{N_m+N-1}\left\{s(n)-\hat{s}(n)\right\}^2}\right) \tag{20}$$

where M represents the number of frames in a signal and N the number of samples per frame. It is well-known that Seg.SNR is more accurate in indicating the speech distortion than the overall SNR. The higher value of the Seg.SNR indicates the weaker speech distortions.

3) Perceptual evaluation of speech quality: PESQ is an objective quality measure algorithm designed to predict the subjective opinion score of a degraded audio sample and it is recommended by ITU-T for speech quality assessment [15]. In PESQ measure, a reference signal and the processed signal are first aligned in both time and level. The PESQ measure was reported to be highly correlated with subjective listening tests in [15] for a large number of testing conditions.

4.2. Subjective Measure-Mean Opinion Score

Subjective measure is based on listener's judgment. In our experimental evaluation, the listening tests have been accomplished with five listeners in a closed room and headphones have been used during experiments. Each listener provides a score between one and five for each test signal. This score represents his overall appreciation of the remnant musical noise and the speech distortion. The scale used for these tests correspond to the MOS


Amplitude
Time (Second)
x 10^4

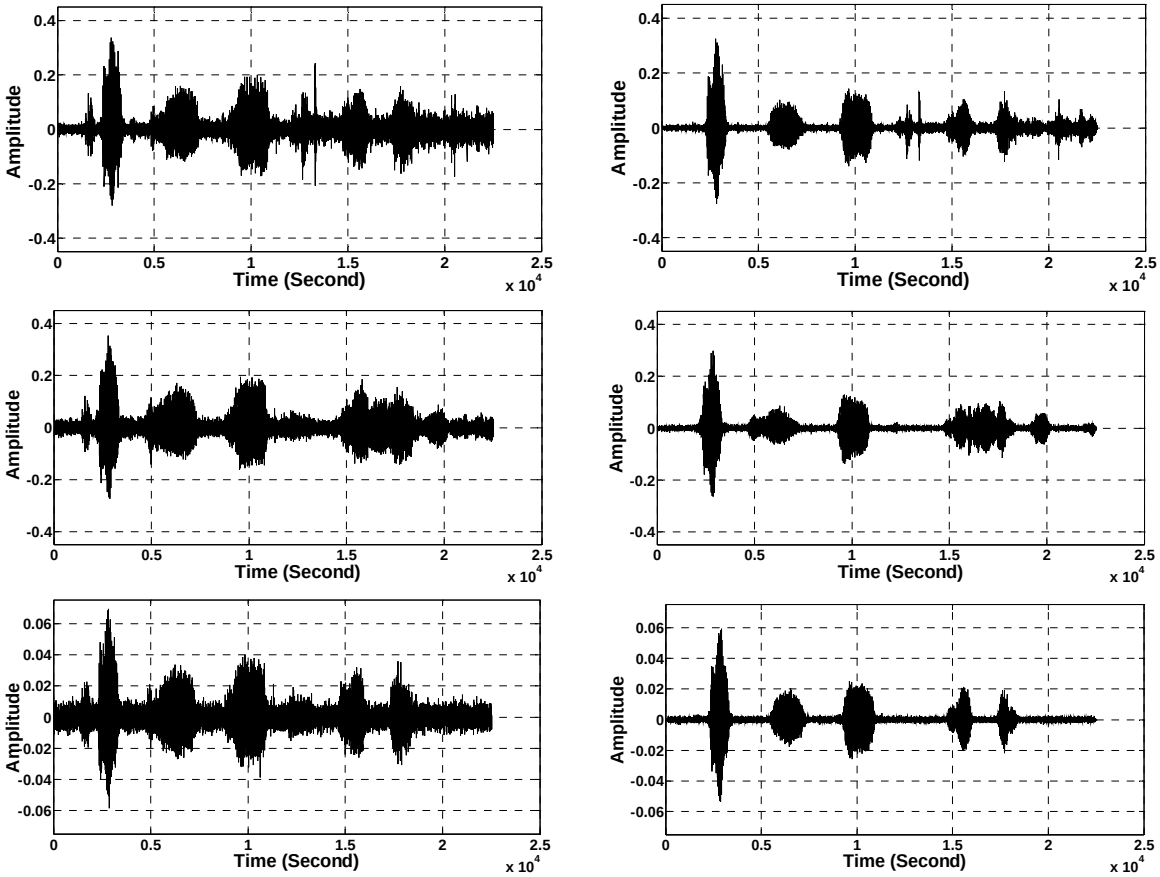


Figure 6. Temporal waveforms of sp1 utterance, "*The birch canoe slid on the smooth planks*", by a male speaker from the NOIZEUS speech corpus: (from top to bottom) clean speech; (left side, from top to bottom) speech degraded by car noise, train noise, babble noise, restaurant noise, airport noise, street noise, exhibition noise, and white noise, respectively (5 dB SNR); (right side, from top to bottom) corresponding enhanced speech.

scale presented in [3]. For each speaker, the following procedure has been applied: 1) clean speech and noisy speech is played and repeated twice; 2) each test signal, which is repeated twice for each score, is played three times in a random order. This leads to 20 scores for each signal.

Table 1 presents the objective evaluation and comparison of the proposed algorithm, IP-MBSS, with MBSS in terms of output SNR values (dB), and output Seg.SNR values (dB) at different labels of SNR. The value of output SNR, and output Seg.SNR for different types of noises for IP-MBSS is observed to be better than MBSS.

The results shown in **Table 2**, presents the PESQ improvement score and MOS score of IP-MBSS over MBSS algorithm. In the case of the PESQ measure, the proposed IP-MBSS technique gives better PESQ scores than the MBSS technique while in MOS case the enhanced speech obtained by proposed algorithm gives poor result for train and airport noise in comparison to MBSS algorithm.

Moreover, speech spectrograms constitute a well-suited tool for observing the remnant noise and speech distortion. It can be seen from **Figures 5-11**, that the musical structure of the remnant noise is reduced more by IP-MBSS, even compared to MBSS. Thus, speech enhanced by the proposed algorithm is more pleasant and the remnant musical noise has a "perceptually white quality" while distortion remains acceptable. This confirms the values of the SNR, Seg.SNR (**Table 1**) and PESQ; also it is validated by listening tests (**Table 2**).

5. Conclusions

In this paper, a multi-band speech enhancement algorithm exploiting iterative processing (IP-MBSS) is proposed for enhancement of speech degraded by non-stationary noises. In the proposed algorithm, IP-MBSS, the output of multi-band spectral subtraction (MBSS) algorithm is used as the input signal again for next iteration process. The iteration is performed to a limited number of times. After the execution of the reference MBSS algorithm, the additive noise changes to remnant musical noise. The remnant noise is re-estimated at each iteration and the spectral over-subtraction is executed separately

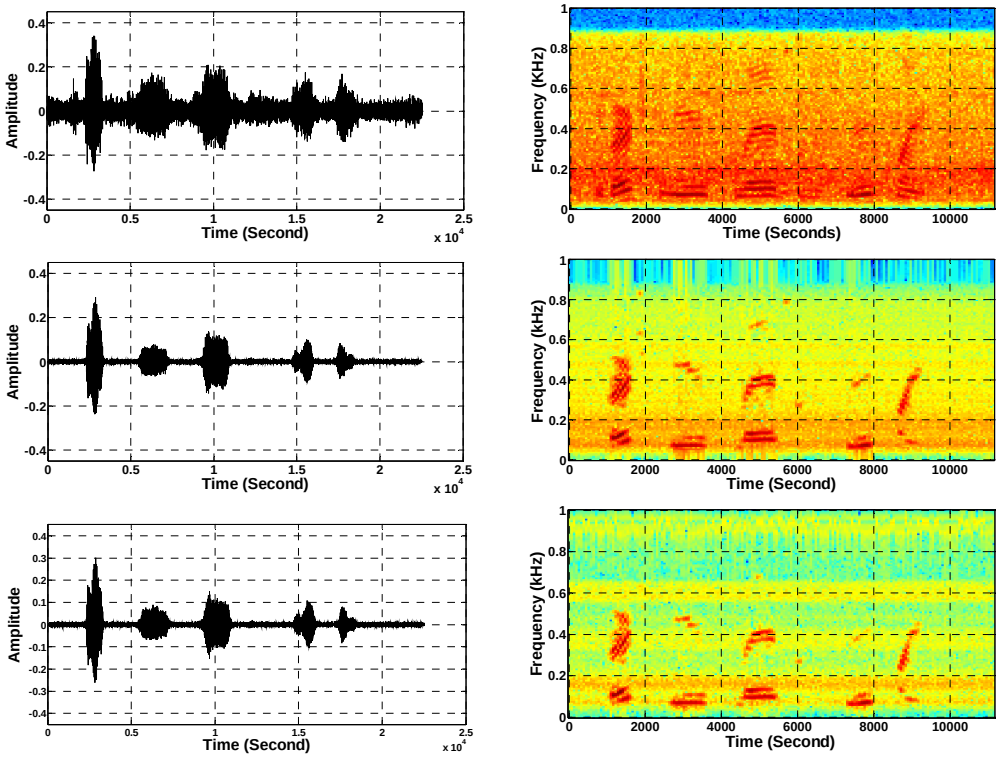


Figure 7. Temporal waveforms and speech spectrogram with sp1 utterance, "*The birch canoe slid on the smooth planks*", by a male speaker from the NOIZEUS speech corpus: (from top to bottom) noisy speech (degraded by car noise at 5 dB SNR); speech enhanced by MBSS (PESQ = 1.776), and speech enhanced by IP-MBSS (1.915).

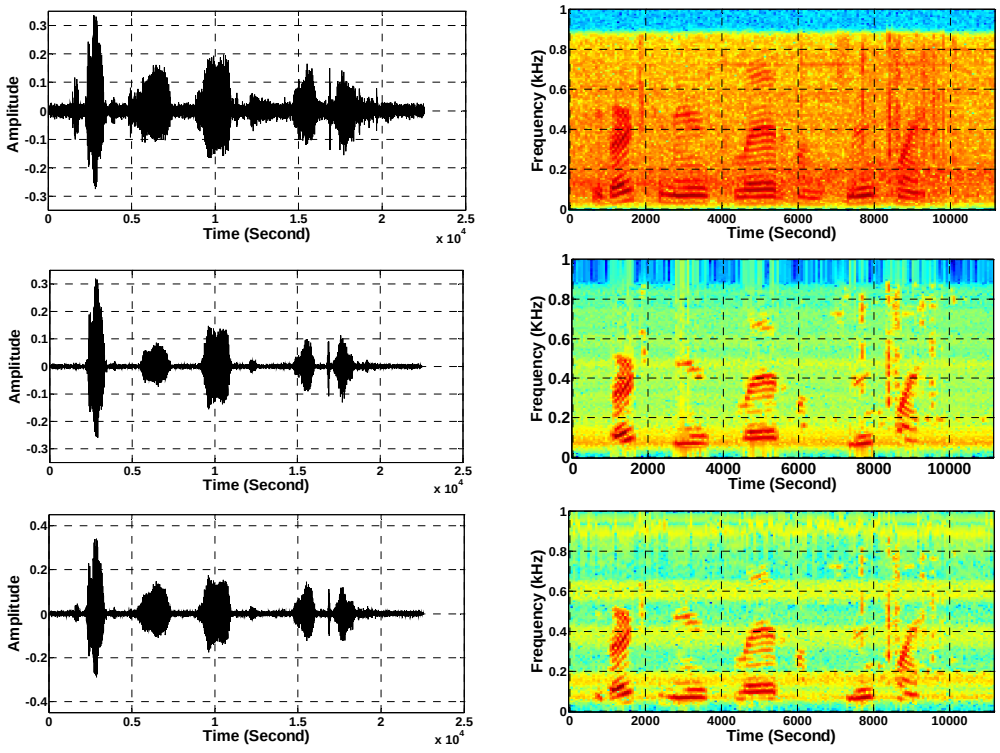


Figure 8. Temporal waveforms and speech spectrogram of sp1 utterance, "*The birch canoe slid on the smooth planks*", by a male speaker from the NOIZEUS speech corpus: (from top to bottom) noisy speech (degraded by car noise at 10 dB SNR); speech enhanced by MBSS (PESQ = 2.030), and speech enhanced by IP-MBSS (PESQ = 2.147).

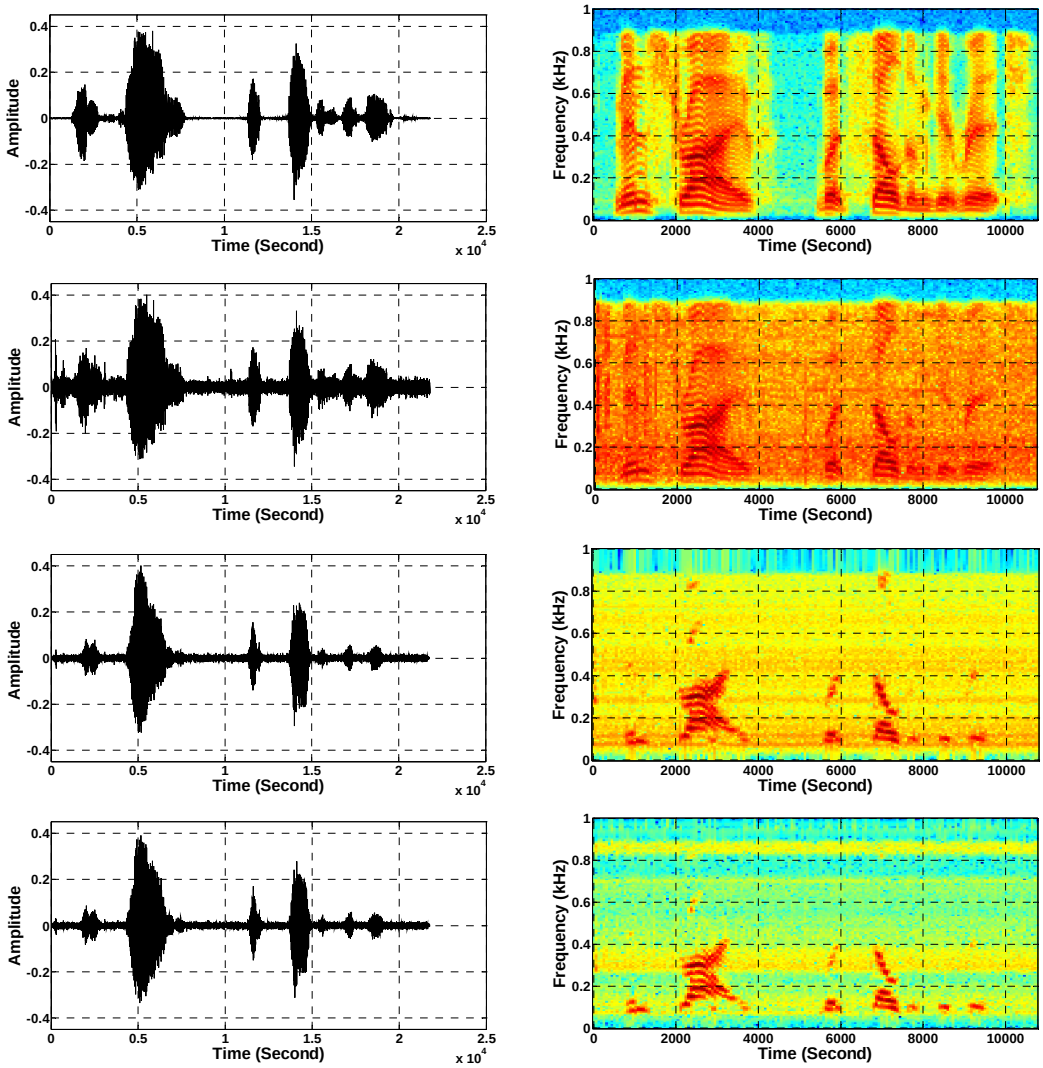
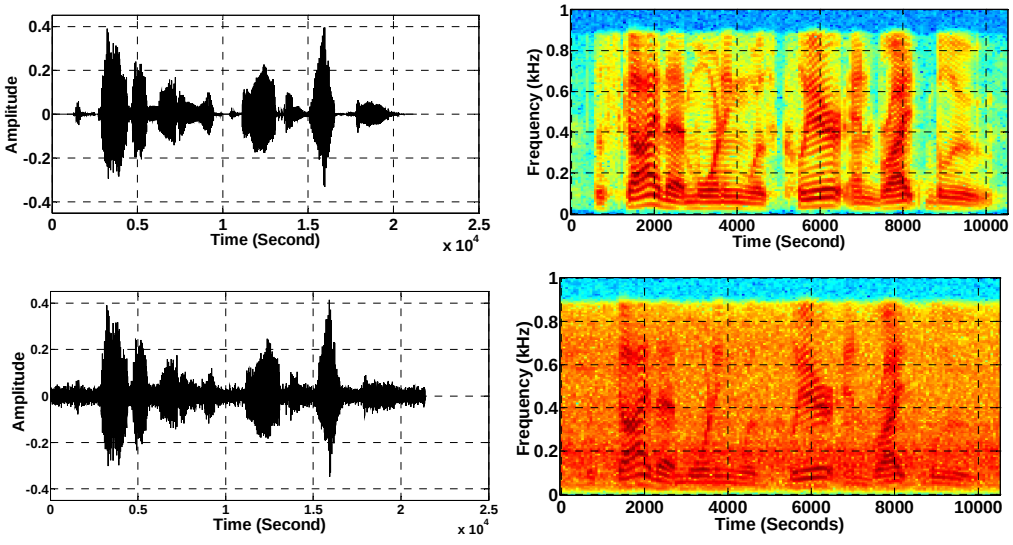


Figure 9. Temporal waveforms and speech spectrograms of sp 6 utterance, *"Men strive but seldom get rich"*, by a male speaker from the NOIZEUS speech corpus: (from top to bottom) clean speech; noisy speech (speech degraded by car noise at 10 dB SNR); speech enhanced by MBSS algorithm (PESQ = 2.157); and speech enhanced by IP-MBSS (PESQ = 2.267).



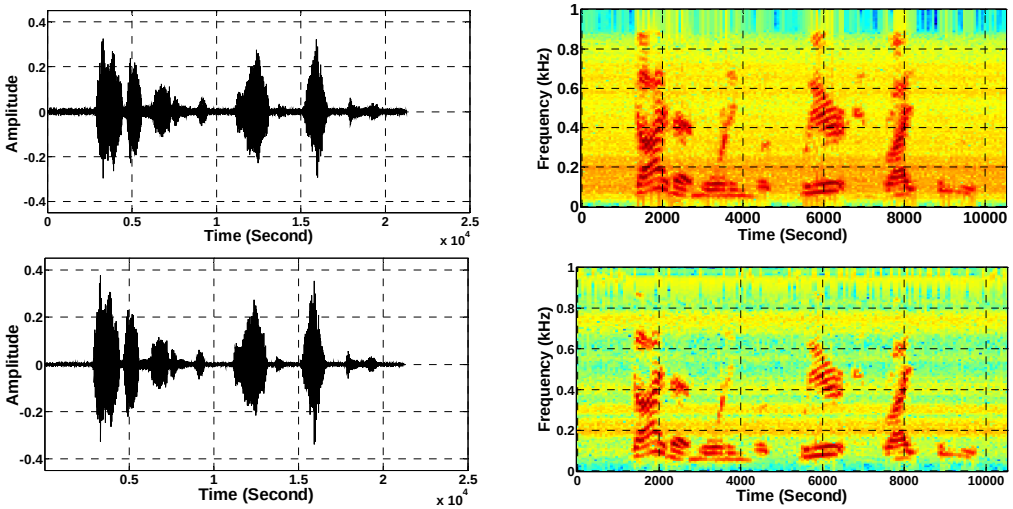


Figure 10. Temporal waveforms and speech spectrograms of sp10 utterance, "*The sky that morning was clear and bright blue*", by a male speaker from the NOIZEUS speech corpus: (from top to bottom) clean speech; noisy speech (speech degraded by car noise at 10 dB SNR); speech enhanced by MBSS (2.259); and speech enhanced by IP-MBSS (2.459).

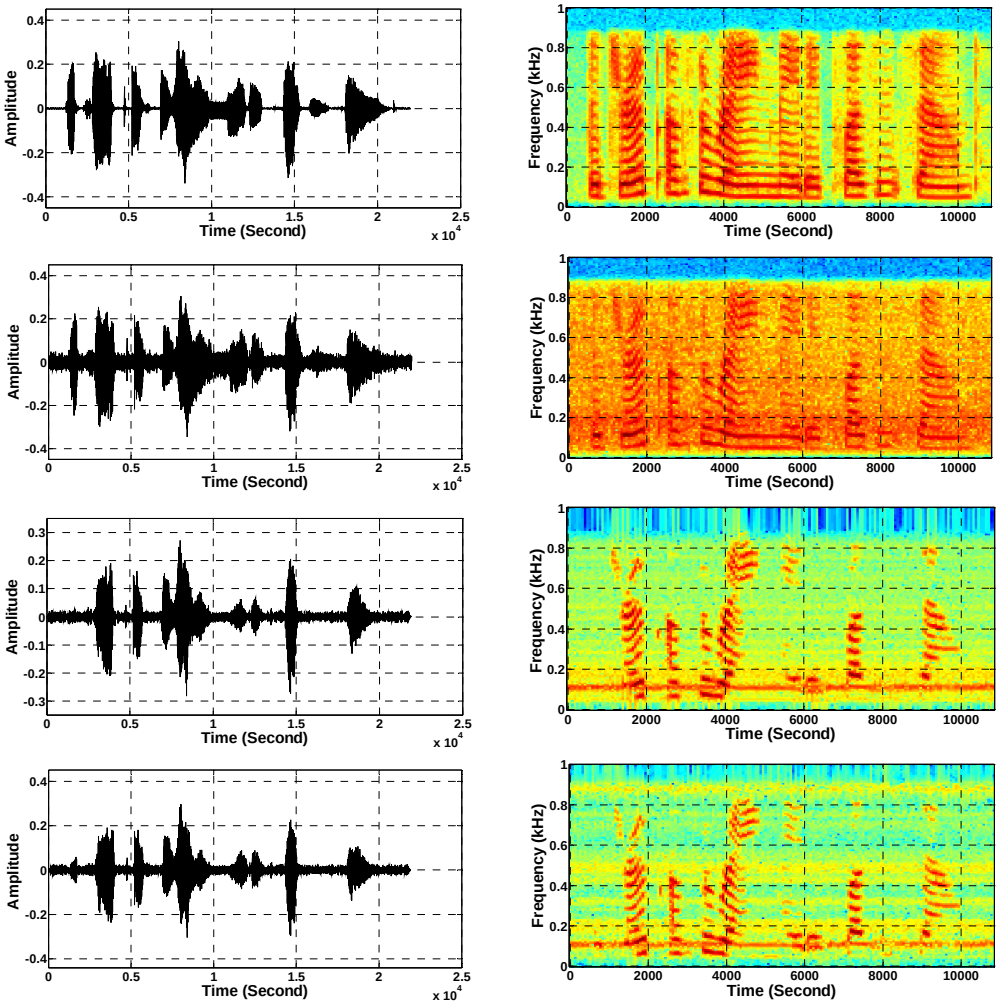


Figure 11. Temporal waveforms and speech spectrograms of sp12 utterance, "*The drip of the rain made a pleasant sound*", by a female speaker from the NOIZEUS speech corpus: (from top to bottom) clean speech; noisy speech (degraded by car noise at 10 dB SNR); speech enhanced by MBSS algorithm (2.005); and speech enhanced by IP-MBSS (2.255).

Table 1. Objective evaluation and comparison of IP-MBSS in terms of output SNR values (dB), and output Seg.SNR values (dB).

Noise Type	Enhancement Algorithms	SNR (dB)				Seg.SNR (dB)			
		0 dB	5 dB	10 dB	15 dB	0 dB	5 dB	10 dB	15 dB
Car	MBSS	4.26	6.01	6.39	6.88	4.19	5.98	6.35	6.82
	IP-MBSS	4.50	6.11	6.46	6.94	4.46	6.10	6.44	6.90
Train	MBSS	3.47	5.82	7.41	7.19	3.42	5.75	7.38	7.17
	IP-MBSS	3.57	5.96	7.33	7.23	3.54	5.92	7.33	7.25
Restaurant	MBSS	2.15	4.60	5.73	6.46	2.10	4.54	5.66	6.45
	IP-MBSS	2.27	5.04	5.84	6.52	2.24	4.99	5.83	6.52
Babble	MBSS	2.27	4.64	6.45	5.92	2.21	4.63	6.42	5.87
	IP-MBSS	2.40	4.89	6.51	5.98	2.35	4.88	6.50	5.97
Airport	MBSS	3.61	4.81	6.26	5.57	3.52	4.76	6.23	5.50
	IP-MBSS	3.71	4.97	6.34	5.68	3.63	4.91	6.33	5.66
Street	MBSS	4.24	5.00	5.68	6.59	4.17	4.89	5.63	6.53
	IP-MBSS	4.42	5.56	5.72	6.66	4.39	5.38	5.68	6.63
Exhibition	MBSS	3.65	3.28	7.12	6.89	3.60	3.20	7.09	6.86
	IP-MBSS	3.92	3.34	7.12	6.91	3.91	3.27	7.11	6.89
White	MBSS	5.09	6.87	7.29	7.49	5.03	6.85	7.28	7.47
	IP-MBSS	5.25	6.86	7.25	7.46	5.23	6.86	7.26	7.46

Table 2. Object and subject evaluation of IP-MBSS in terms of PESQ and MOS score.

Noise Type	Enhancement Algorithms	PESQ Improvement Score				MOS Score			
		0 dB	5 dB	10 dB	15 dB	0 dB	5 dB	10 dB	15 dB
Car	MBSS	1.615	1.776	2.030	2.293	1.8	2.7	3.5	4.3
	IP-MBSS	1.693	1.915	2.147	2.489	2	2.8	3.6	4.1
Train	MBSS	1.608	1.886	1.850	2.166	2.6	3.3	3.7	4.2
	IP-MBSS	1.693	1.893	2.010	2.353	2.3	2.9	3.4	4.2
Restaurant	MBSS	1.697	1.885	2.039	2.295	1.8	2.7	3.5	4.0
	IP-MBSS	1.787	1.927	2.187	2.479	1.9	2.7	3.4	4.1
Babble	MBSS	1.665	1.907	2.134	2.237	1.6	2.7	3.6	4.2
	IP-MBSS	1.667	2.036	2.341	2.413	1.8	2.7	3.6	4.3
Airport	MBSS	1.774	1.953	2.161	2.263	1.8	2.8	3.6	4.2
	IP-MBSS	1.876	2.061	2.294	2.471	1.6	2.1	2.8	3.9
Street	MBSS	1.416	1.866	2.002	2.300	1.8	2.6	3.5	4.2
	IP-MBSS	1.614	1.956	2.190	2.501	2	2.7	3.5	4.2
Exhibition	MBSS	1.298	1.633	2.001	2.260	1.8	2.7	3.4	4
	IP-MBSS	1.379	1.782	2.102	2.420	1.9	2.6	3.8	4.4
White	MBSS	1.433	1.669	2.069	2.297	2.6	3.5	4.1	4.5
	IP-MBSS	1.602	1.901	2.235	2.474	2.9	3.6	4	4.4

in each band. A comparison with the reference MBSS algorithm is carried out to evaluate the performance of the proposed enhancement algorithm.

Furthermore, the simulations results, with different types of noises, have shown that the proposed algorithm, IP-MBSS, with appropriate iteration number reduces the remnant musical noise tones efficiently that appear in the case of MBSS algorithm and improves the quality and intelligibility of the enhanced speech. Therefore, the performance gain of IP-MBSS, in comparison to MBSS, is found to be more pronounced for the case of low SNRs. It is also evident from the subjective listening tests that

the speech enhanced by IP-MBSS algorithm does contains a little amount of remnant noise and speech distortion. The remnant noise is of perceptually white quality and the distortions stay within acceptable limit.

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 IEEE*, Vol. 80, No. 10, 1992, pp. 1526-1555. doi:10.1109/5.168664

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

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

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

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

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

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

[9] S. Kamath and P. Loizou, "A Multi-Band Spectral Sub-Traction Method for Enhancing Speech Corrupted by Colored Noise," *Proceedings of International Conference on Acoustics*, *Speech*, *and Signal Processing*, Orlando, 13-17 May 2002, pp. IV-4164. doi:10.1109/ICASSP.2002.5745591

[10] N. Upadhyay and A. Karmakar, "The Spectral Subtractive-Type Algorithms for Enhancement of Noisy Speech: A Review," *International Journal of Research and Reviews in Signal Acquisition and Processing*, Vol. 1, No. 3, 2011, pp. 43-49.

[11] N. Upadhyay and A. Karmakar, "Single Channel Speech Enhancement Utilizing Iterative Processing of Multi-Band Spectral Subtraction Algorithm," *Proceedings of International Conference on Power*, *Control and Embedded Systems*, Allahabad, 17-19 December 2012, pp. 196-201.

[12] S. Ogata and T. Shimamura, "Reinforced Spectral Subtraction Method to Enhance Speech Signal," *Proceedings* of *International Conference on Electrical and Electronic Technology*, Vol. 1, 2011, pp. 242-245.

[13] S. Li, J.-Q. Wang, M. Niu, X.-J. Jing and T. Liu, "Iterative Spectral Subtraction Method for Mil-Limeter Wave Conducted Speech Enhancement," *Journal Biomedical Science and Engineering*, Vol. 3, No. 2, 2010, pp. 187-192. doi:10.4236/jbise.2010.32024

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

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

