I.J. Image, Graphics and Signal Processing, 2013, 11, 13-22
Published Online September 2013 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2013.11.02

Spectral Subtractive-Type Algorithms for Enhancement of Noisy Speech: An Integrative Review

Navneet Upadhyay[1], Abhijit Karmakar[2]
[1]Department of Electrical & Electronics Engineering, Birla Institute of Technology and Science, Pilani 333031, India
[2]Integrated Circuit Design Group, CSIR - Central Electronics Engineering Research Institute, Pilani 333031, India
e-mail: navneet_upd@rediffmail.com[1], abhijit @ceeri.ernet.in[2]

Abstract — The spectral subtraction method is a classical approach for enhancement of speech degraded by additive background noise. The basic principle of this method is to estimate the short-time spectral magnitude of speech by subtracting estimated noise spectrum from the noisy speech spectrum. This is also achieved by multiplying the noisy speech spectrum with a gain function and later combining it with the phase of the noisy speech. Besides reducing the background noise, this method introduces an annoying perceptible tonal characteristic in the enhanced speech and affects the human listening, known as remnant musical noise. Several variations and implementations of this method have been adopted in past decades to address the limitations of spectral subtraction method. These variations constitute a family of subtractive-type algorithms and operate in frequency domain. The objective of this paper is to provide an extensive overview of spectral subtractive-type algorithms for enhancement of noisy speech. After the review, this paper is concluded by mentioning a future direction of speech enhancement research from spectral subtraction perspective.

Index Terms — Speech enhancement, additive background noise, noise estimation, spectral subtractive-type algorithms, remnant musical noise

I. INTRODUCTION

Speech is one of the most prominent and primary modes of interaction between human-to-human and human-to-machine communication in various fields for instance automatic speech recognition and speaker identification [1]. The present day speech communication systems are severely degraded due to various types of unwanted random sound which make the listening task difficult for a direct listener and cause inaccurate transfer of information [2]. Therefore, enhancement speech is one of the main motives of various researching endeavors in the field of speech processing over the past few decades. The main objective of speech enhancement 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 quality and/or intelligibility.

The quality of speech is a subjective measure which reflects the way that the signal is perceived by listeners. Intelligibility, on the other hand is an objective measure of the amount of information that can be extracted by listeners from the speech signal. These two measures are uncorrelated and independent of each other. A speech signal may be of high quality and low intelligibility and vice-versa [1-4].

The classification of speech enhancement method depends on the number of microphones that are used for collecting speech data, into single, dual or multi-channel. Although the multi-channel speech enhancement is better than that of single channel speech enhancement [1-2], yet the single channel speech enhancement is still a significant field of researching because of its simple implementation and ease of computation. Single channel speech enhancement uses only one microphone to collect noisy speech data [1-4].

The estimation of the spectral amplitude from the noisy data is easier than estimate of both the amplitude and phase. In [5-6], revealed that the short-time spectral amplitude (STSA) is more important than the phase information for the quality and intelligibility of speech. Therefore, single channel speech enhancement is usually divided into two classes based on the STSA estimation. The first class applies subtractive-type algorithms and attempt to estimate the short-time spectral magnitude (STSM) of speech by subtracting the estimated noise spectrum. Here, noise is estimated during speech pauses [7-11, 13]. The other class applies a spectral subtraction filter (SSF) to the noisy speech, so that the spectral amplitude of enhanced speech can be obtained. The design principle is to select appropriate parameters of the filter to minimize the difference between the enhanced speech and the clean speech signal [8].

In real-world listening environments, the speech is mostly degraded by additive noises [5, 9-14]. Additive noise is typically background noise which is uncorrelated with the clean speech signal in nature like white Gaussian noise (WGN), colored noise, multi-talker (babble) noise. The background noise may be stationary or non-stationary in nature. Therefore, the

noisy signal can be modeled as a sum of the clean speech and the noise signal [9-11, 13] as

$$y(n) = s(n) + d(n),\ \ n = 0, 1, 2, .., (N - 1) \qquad (1)$$

where n is the discrete-time index, and N is the number of samples in the signal. Also, $y(n)$, $s(n)$, and $d(n)$, are the n^{th} sample of the discrete-time signal of noisy speech, clean speech and random noise, reactively. Although speech is non-stationary in nature whose spectral properties vary with time, usually the short-time Fourier transform (STFT) is used to divide the speech signal in small frames for further processing [9-15]. 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 [9 -15],

$$Y_W(\omega) = S_W(\omega) + D_W(\omega) \qquad (2)$$

where ω is the discrete-frequency index of the frame and W is the window (Hamming or Henning window). Throughout this paper, it is assumed that the signal is segmented into frames first and then windowed, hence for simplicity, we drop the use of subscript W from windowed signals. For implementation of speech enhancement method, few assumptions are necessary. First, the speech signal should be stationary; and other, the noise is assumed to be zero mean and uncorrelated with clean speech signal.

The goal of this paper is to provide an integrative review of subtractive-type noisy speech enhancement algorithms. In addition to basic spectral subtraction algorithm [7, 9, 10] other most notable algorithms are spectral over-subtraction (SOS) [15], parametric spectral subtraction (PSS) [16], spectral subtraction based on cross correlation [17], non-linear spectral subtraction (NSS) [18], multi-band spectral subtraction (MBSS) [19], Wiener filtering (WF) [20], iterative spectral subtraction (ISS) [21], extended spectral subtraction (ESS) [22], and spectral subtraction based on perceptual properties (SSPP) [23].

This paper is organized as follows; we start with historical account on the use of enhancement methods of noisy speech. In section II, the principle of spectral subtraction method has been presented. Section III presents various modified forms of subtractive-type algorithms. Finally, the conclusion of review has been provided in section IV.

II. Principle of Spectral Subtraction Method

The spectral subtraction is one of the most well-known and computationally efficient 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, 9, 10] is based on non-parametric approach, which simply needs an estimate of noise spectrum and used for both speech enhancement and speech recognition. The spectral subtraction method mainly, involves two phases. In the first phase the average estimate of the noise spectrum is subtracted from the instantaneous spectrum of the noisy speech. This is termed as basic spectral subtraction (BSS) step. In the second phase, 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. It is assumed that the phase of noise has no effect on phase of clean speech because change of phase in the process is not perceived by human ear [5, 6]. Therefore, STSM of noisy speech is equal to the sum of STSM of clean speech and STSM of random noise without the phase information and (2) can be expressed [11] as

$$|Y(\omega)| = |S(\omega)| + |D(\omega)| \qquad (3)$$

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

$$|Y(\omega)|^2 = |S(\omega)|^2 + |D(\omega)|^2 + S(\omega)D^*(\omega) + S^*(\omega)D(\omega) \qquad (4)$$

Here, $D^*(\omega)$ and $S^*(\omega)$ are the complex conjugates of $D(\omega)$ and $S(\omega)$, respectively. The $|Y(\omega)|^2$, $|S(\omega)|^2$, and $|D(\omega)|^2$, are referred to as the short-time spectrum of noisy speech, clean speech and random noise, respectively. The value of $|D(\omega)|^2$, $S(\omega)D^*(\omega)$ and $S^*(\omega)D(\omega)$ cannot be obtained directly and are approximated as, $E\{|D(\omega)|^2\}$, $E\{S(\omega)D^*(\omega)\}$ and $E\{S^*(\omega)D(\omega)\}$,where $E\{.\}$ denotes the ensemble averaging operator. As the additive noise assumed to be zero mean and uncorrelated with the clean speech signal, the terms $E\{S(\omega)D^*(\omega)\}$ and $E\{S^*(\omega)D(\omega)\}$ reduce to zero [11, 13]. Therefore, (4) can be rewritten as

$$|Y(\omega)|^2 = |S(\omega)|^2 + |D(\omega)|^2 \qquad (5)$$

It is desired to choose an estimate $|\hat{S}(\omega)|$ that will minimize the error

$$E_w(\omega) = \left||\hat{S}(\omega)|^2 - |S(\omega)|^2\right|, \qquad (6)$$

$$E_w(\omega) = ||\hat{S}(\omega)|^2 - |Y(\omega)|^2 + E\{|D(\omega)|^2\}|, \qquad (7)$$

The (7) can be minimized by choosing

$$|\hat{S}(\omega)|^2 = |Y(\omega)|^2 - |\widehat{D}(\omega)|^2 \qquad (8)$$

where $|\hat{S}(\omega)|^2$ is the short-time spectrum of estimated speech and $|\widehat{D}(\omega)|^2$ is the average noise power which normally, estimated and updated during speech pauses.

In spectral subtraction method, it is assumed that the speech signal is degraded by additive white Gaussian noise (AWGN) and the spectrum of white noise is flat. Hence, the noise affects the speech signal uniformly

over the whole spectrum. In this method, the subtraction process needs to be done 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 power spectrum of estimated speech can become negative due to over-estimation of noise, but to get rid of this possibility, therefore, a HWR (by setting the negative portions to zero) or full-wave rectification (absolute value) are introduced. But the HWR introduces annoying noise in the enhanced speech. Whereas, full-wave rectification (FWR) avoids the creation of annoying noise, but it is less effective in suppressing noise. Therefore, HWR is often used in spectral subtraction method due to its superior noise suppression ability. Thus, the complete power spectral subtraction algorithm is given by

$$|\hat{S}(\omega)|^2 = \begin{cases} \{|Y(\omega)|^2 - |\widehat{D}(\omega)|^2\}, \\ \qquad \text{if } |Y(\omega)|^2 > |\widehat{D}(\omega)|^2 \\ 0 \qquad \text{else} \end{cases} \tag{9}$$

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 [11-14], can be expressed as

$$\hat{s}(n) = \text{ ISTFT}\{|\hat{S}(\omega)|.\exp(j\varphi_y(\omega))\} \tag{10}$$

On the contrary, a generalize form of spectral subtraction (8) can be obtained by altering the power exponent from 2 to b, which determines the sharpness of the transition.

$$|\hat{S}(\omega)|^b = |Y(\omega)|^b - |\widehat{D}(\omega)|^b, b > 0 \tag{11}$$

where $b = 2$ represents the power spectral subtraction and $b = 1$ represents the magnitude spectral subtraction. Figure 1, shows the block diagram of spectral subtraction method.

A. Noise Estimation

The noise estimation is the most critical part of frequency domain enhancement algorithms because the quality of the enhanced speech depends on the accurate noise spectrum estimation [11].

The noisy signal consists of some portions that have speech activities and some portions that have non-speech activity called speech pauses. The speech activities means that the portions of noisy speech consists of speech, which is degraded by background noise, whereas the speech pauses are the parts of the noisy speech only with background noise. Moreover, the speech regions are periodic in nature and energy of speech regions is larger than that of non-speech regions while non-speech sounds are more noise-like and have more energy than silence. Silence has the least amount of energy and is the representation of the background noise of the environment. As a result, the SNR's of speech activity regions are generally higher than that of non-speech regions. Therefore, the enhancement of speech regions is more effective than that in speech pauses [11].

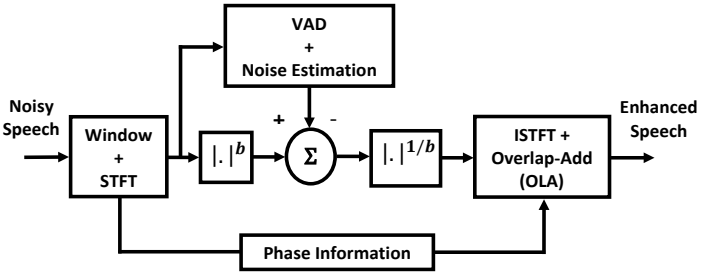

Figure 1. Block diagram of spectral subtraction method.

In spectral subtraction method, voice activity detection (VAD) algorithm plays a central role for detecting presence and absence of speech in a noisy speech signal. It gives the values of zeros and one as an indicator of speech pauses and speech activity in each frame. If the noise is stationary, the first 100-200 ms of noisy signal is assumed to be pure noise. To do this, a good estimation can be resulted by computing the average of the noise in silence frame spectra [11].

In presence of non-stationary noise, the noise spectrum needs to be estimated and updated, continuously. The noise estimation can be updated by using the first order relation as

$$|\widehat{D}(\omega,k)|^2 = \lambda|\widehat{D}(\omega,k-1)|^2 + (1-\lambda)|Y(\omega,k)|^2 \tag{12}$$

where λ ($0 \leq \lambda \leq 1$) is a time and frequency dependent smoothing parameter whose value depends on the noise changing rate and k refers to the current frame-index. $|Y(\omega,k)|^2$ represent the short-time power spectrum of noisy speech, $|\widehat{D}(\omega,k)|^2$ is the updated noise spectral estimate, and $|\widehat{D}(\omega,k-1)|^2$ is the past noise spectral estimate [18].

In [24, 25] suggested algorithms are based on finding the minimum statistics of noisy speech for each sub band over a time window.

B. Limitation of Spectral Subtraction Algorithm

The major weakness of spectral subtraction method is that after the processing, the enhanced speech is accompanied by excessive remnant noise with musical nature. As a result, the detection of speech pauses is difficult. This noise is generated due to the in-accurate estimation of noise from each frame i.e. mismatch between the noise spectrum estimate and the instantaneous noise spectrum [11]. This noise sometimes more disturbing not only for human ear, but also for speaker recognition systems. Several publications have been existed in the literature for the modifications of the spectral subtraction method to combat the problem of remnant noise and musical noise artifacts.

III. Spectral Subtractive-Type Algorithms

A. Spectral Over-Subtraction Algorithm

An improved version of spectral subtraction method was proposed in [15] to minimize the annoying musical noise. In this algorithm, the spectral subtraction method [9] is used by using two additional parameters, over-subtraction factor α, and spectral floor parameter β [15]. The algorithm [15] can be described as

$$|\hat{S}(\omega)|^2 = \begin{cases} |Y(\omega)|^2 - \alpha.|\hat{D}(\omega)|^2, & \text{if } \frac{|\hat{D}(\omega)|^2}{|Y(\omega)|^2} < \frac{1}{\alpha+\beta} \\ \beta.|\hat{D}(\omega)|^2, & \text{else} \end{cases}$$

$$\text{with} \quad \alpha \geq 1 \text{ and } 0 \leq \beta \ll 1 \tag{13}$$

The function of over-subtraction factor is to control the amount of noise power spectrum subtracted from the noisy speech power spectrum and the introduction of spectral floor parameter prevents the spectral components of the resultant spectrum to fall below a preset minimum level rather than setting to zero. To reduce the speech distortion caused by large value of α, its value is adapted from frame to frame. The basic idea is take into account that the subtraction process must depends on segmental SNR. Therefore, the over-subtraction factor can be calculated as

$$\alpha = \alpha_0 + (\text{SNR} - \text{SNR}_{\text{min}}) \left(\frac{\alpha_{\text{min}} - \alpha_0}{\text{SNR}_{\text{max}} - \text{SNR}_{\text{min}}} \right),$$

$$\text{SNR}_{\text{min}} \leq \text{SNR} \leq \text{SNR}_{\text{max}} \tag{14}$$

where

$$\text{SNR (dB)} = 10 \log_{10} \left(\frac{\sum_{k=0}^{N-1} |Y(\omega)|^2}{\sum_{k=0}^{N-1} |\hat{D}(\omega)|^2} \right) \tag{15}$$

Here, the value of $\alpha_{\text{min}} = 1$, $\alpha_{\text{max}} = \alpha_0$, $\text{SNR}_{\text{min}} = 0$ dB, $\text{SNR}_{\text{max}} = 20$ dB and α_0 ($\alpha_0 \approx 4$), used in (14), 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 segmental SNR is shown in Figure 2.

This implementation assumes that the noise affects the speech spectrum uniformly and the performance of this scheme is restricted in the usage of fixed value of subtraction parameters, which are difficult for real-world noises. Thus, it is not easy to reduce noise without decreasing speech intelligibility and distortion, especially at very low SNRs. In Figure 3, the block diagram of spectral over-subtraction algorithm is shown.

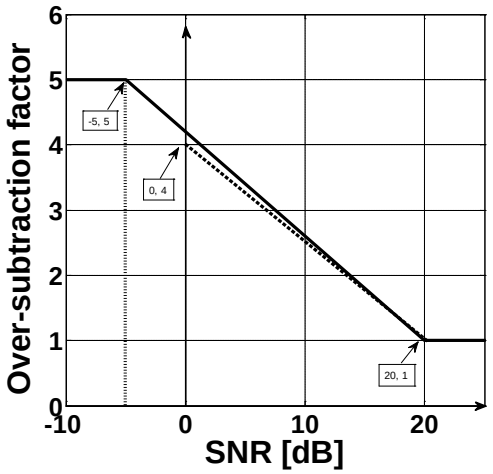

Figure 2. The relation between over-subtraction factor and segmental SNR.

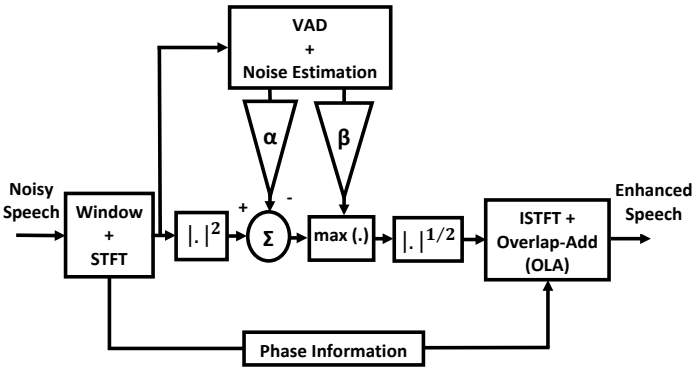

Figure 3. Block diagram of spectral over-subtraction algorithm.

B. Parametric Spectral Subtraction Algorithm

In [15], the subtractive parameters have been computed experimentally and have not been selected optimally in any sense. An algorithm is proposed in [16], where the subtractive parameters are selected in mean squared error (MSE) sense to reduce the remnant noise problem linked with spectral subtraction algorithm. The values of subtractive parameters are derived by the parametric formulation of generalized spectral subtraction algorithm (11). The generalized form of spectral subtraction algorithm can also be given as

$$|\hat{S}(\omega)|^b = a(\omega).|Y(\omega)|^b - b(\omega).|\hat{D}(\omega)|^b \tag{16}$$

where $a(\omega)$ and $b(\omega)$ are the algorithm parameters. In parametric spectral subtraction algorithm [16], considering $a(\omega) = b(\omega)$ and derived a different estimator as

$$|\hat{S}(\omega)| = \left\{ \frac{\varepsilon^b(\omega)}{\delta_b + \varepsilon^b(\omega)} . \left[|Y(\omega)|^b - |\hat{D}(\omega)|^b \right] \right\}^{1/b} \tag{17}$$

Here $\varepsilon(\omega)$ is the a-*priori* SNR and δ_b is constant for a given power exponent b. The values of δ_b are 0.2146, 0.5 and 0.7055 for b = 1, 2 and 3, respectively [11]. The value of $\varepsilon(\omega)$ cannot be computed exactly as

we do not have access to the clean speech signal. In [8], proposed the approach and in [16], approximated as

$$\varepsilon(\omega) = (1-\eta).\left(\frac{|S(\omega)|^2_{\text{current}}}{|\hat{D}(\omega)|^2_{\text{current}}}\right) + \eta.\frac{|\hat{S}(\omega)|^2_{\text{previous}}}{|\hat{D}(\omega)|^2_{\text{previous}}} \quad (18)$$

Here $|S(\omega)|^2_{\text{current}} = \max\left(|Y(\omega)|^2 - |\hat{D}(\omega)|^2, 0\right)$ and η is a smoothing constant also $|\hat{S}(\omega)|^2_{\text{previous}}$ is the enhanced spectrum computed in previous frame. The first term of (18) is the value of current SNR and the second term is previous SNR. Also, the spectral floor with lower bound $\overline{\mu.Y}$ is used to limit the signal attenuation. The final constrained parametric estimator is implemented as

$$\bar{S}(\omega) = \begin{cases} \left|\hat{S}(\omega)\right|, & if\ |\hat{S}(\omega)| \geq \mu.|Y(\omega)| \\ \overline{\mu.Y}, & \text{else} \end{cases} \quad (19)$$

where μ is the spectral flooring constant $0 < \mu < 1$, and $\bar{S}(\omega)$ is the previous estimated amplitude.

C. Spectral Subtraction based on Cross Correlation

The spectral subtraction based on cross correlation was proposed in [17]. In this algorithm, it is assumed that the noise signal is correlated with clean speech signal. Therefore, in (4) the cross terms cannot be ignore, these terms are used to represent the cross correlations between clean speech and correlated noise. Also, we do not have access to clean speech signal so it is very difficult to estimate cross correlations between clean speech signal and correlated noise. But, since we have access to the noisy speech signal, we can get an estimate of the cross correlation by computing the cross correlation between noisy speech and noise signal.

D. Non-linear Spectral Subtraction

The non-linear spectral subtraction (NSS) algorithm, proposed in [18], is motivated by the fact that algorithms with fixed subtraction parameters are unable to adapt well to the varying noise levels and characteristics. This approach is a basically a modification on the algorithm proposed in [15] by making the over-subtraction factor frequency dependent and the subtraction process non-linear. Larger values are subtracted at lower SNR and at higher SNR the subtraction applied is minimal. The NSS algorithm can be written as follows:

$$|\hat{S}(\omega)| = \begin{cases} \overline{|Y(\omega)|} - \phi(\omega), \\ \quad \text{if } \overline{|Y(\omega)|} > \phi(\omega) + \beta.|\hat{D}(\omega)| \\ \beta.\overline{|Y(\omega)|}, \quad \text{otherwise} \end{cases} \quad (20)$$

where β is the spectral floor parameter [15], $\overline{|Y(\omega)|}$ and $|\hat{D}(\omega)|$ the smoothed estimates of noisy speech and noise, respectively. The $\phi(\omega)$ is a non-linear function, calculated for each frame and is dependent on the following parameters:

$$\phi(\omega) = f\left\{\alpha(\omega), \rho(\omega), |\hat{D}(\omega)|\right\} \quad (21)$$

Here, the over-subtraction factor $\alpha(\omega)$ is computed for each frame k as the maximum noise spectrum (estimated during speech pauses) over the last 40 frames:

$$\alpha(\omega) = \max\left(|\hat{D}_k(\omega)|\right)_{k-40 \leq j \leq k} \quad (22)$$

$\rho(\omega)$ is the *a-posteriori* SNR, and is estimated according to the following relation

$$\rho(\omega) = \frac{\overline{|Y(\omega)|}_\rho}{|\hat{D}(\omega)|} \quad (23)$$

where $\overline{|Y(\omega)|}_\rho$ is the noisy speech spectrum smoothed, with a time-frequency dependent smoothing parameter of value 0.5.

E. Multi-band Spectral Subtraction Algorithm

In real-world environment, the noise spectrum is non-uniform over the entire spectrum. Some of the frequencies are affected more adversely than others, depending on the spectral characteristics of the noise, which eventually mean that this kind of noise is non-stationary or colored. To take into account the fact that colored noise affects the speech spectrum differently at different frequencies, a multi-band uniformly spaced frequency approach to spectral over-subtraction [15] is presented in [19].

The result of an implementation of four uniformly spaced frequency bands [19] with estimated segmental SNR of 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)} of noisy speech spectrum is shown in Figure 4. It can be seen from the figure that the segmental SNR of the low frequency bands (Band 1) is significantly higher than the segmental SNR of the high frequency bands (Band 4) [11, 19]. This phenomenon suggests that the noise signal does not affect the speech signal uniformly over the whole spectrum; therefore, subtracting a constant factor of noise spectrum over the whole frequency spectrum may remove speech also.

The multi-band spectral subtraction algorithm [19] is the case of NSS [16]. In this algorithm, the noisy speech spectrum is divided into four uniformly spaced non-overlapping frequency bands, and spectral over-subtraction is performed in each band, separately. This algorithm re-adjusts the over-subtraction factor in each band. Therefore, the estimate of the clean speech spectrum in the i^{th} Band is obtained by

$$|\hat{S}_i(\omega)|^2 = \begin{cases} |Y_i(\omega)|^2 - \alpha_i.\delta_i.|\hat{D}_i(\omega)|^2, \\ \quad \text{if } |\hat{S}_i(\omega)|^2 > \beta.|Y_i(\omega)|^2 \\ \beta.|Y_i(\omega)|^2, \quad \text{else} \end{cases}$$
$$\text{where} \quad \omega_i < \omega < \omega_{i+1} \quad (24)$$

where ω_i and ω_{i+1} are the start and end frequency bins of the i^{th} frequency band, α_i is the band specific over-

subtraction factor which is the function of the segmental SNR of corresponding band. The segmental SNR of i^{th} Band can be computed as

$$\text{SNR}_i\ (\text{dB}) = 10\log_{10}\left(\frac{\sum_{\omega=\omega_i}^{\omega_{i+1}}|Y_i(\omega)|^2}{\sum_{\omega=\omega_i}^{\omega_{i+1}}|\widehat{D}_i(\omega)|^2}\right) \quad (25)$$

The band specific over-subtraction factor can be calculated, using Figure 2, as

$$\alpha_i = \begin{cases} \alpha_{\max}, & \text{if } \text{SNR}_i \le \text{SNR}_{\min} \\ \alpha_{\max} + (\text{SNR}_i - \text{SNR}_{\min})\left(\frac{\alpha_{\min}-\alpha_{\max}}{\text{SNR}_{\max}-\text{SNR}_{\min}}\right), & \text{if } \text{SNR}_{\min} \le \text{SNR}_i \le \text{SNR}_{\max} \\ \alpha_{\min}, & \text{if } \text{SNR}_i \ge \text{SNR}_{\max} \end{cases} \quad (26)$$

Here $\alpha_{\min} = 1, \alpha_{\max} = 5, \text{SNR}_{\min} = -5\text{ dB}, \text{SNR}_{\max} = 20\text{ dB}$.

The δ_i is an additional band subtraction factor that provide an additional degree of control within each band. The values of δ_i used in [19] is empirically calculated as most of the speech energy is concentrated below 1 kHz. The negative values of the estimated spectrum are floored.

As the real-world noise is highly random in nature. So, improvement in the MBSS algorithm for reduction of WGN is required. However, the performance of MBSS method is better than other subtractive-type algorithm. This algorithm has been applied in different configuration in [26-31]. In [29], perceptually motivated un-decimated wavelet packet filterbank is used to obtain bands. In Figure 5, the block diagram of multi-band spectral subtraction algorithm is shown.

F. Wiener Filtering

The spectral subtraction method [9] can also be viewed as a filtering operation [5, 8]. The noisy speech is filtered with a time-variant linear filter where high SNR regions of the measured spectrum are attenuated less than low SNR regions. Therefore, (11) can be expressed as the product of noisy speech spectrum and a spectral subtraction filter (SSF) as

$$|\widehat{S}(\omega)|^b = |Y(\omega)|^b - |\widehat{D}(\omega)|^b = H(\omega).|Y(\omega)|^b \quad (27)$$

$$\text{where} \qquad H(\omega) = \left[1 - \frac{|\widehat{D}(\omega)|^b}{|Y(\omega)|^b}\right] \quad (28)$$

The $H(\omega)$ is a real function, called the gain of SSF. The gain of SSF has zero phase and its magnitude

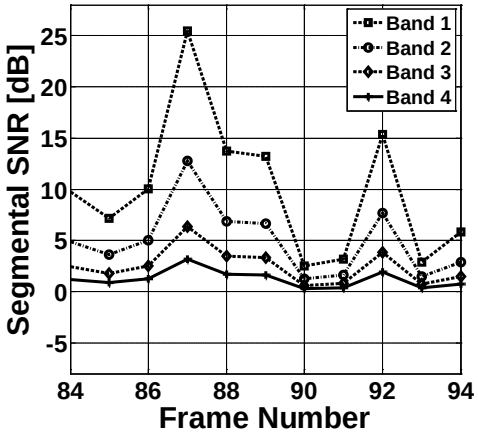

Figure 4. The segmental SNR of four uniformly spaced frequency bands of degraded speech.

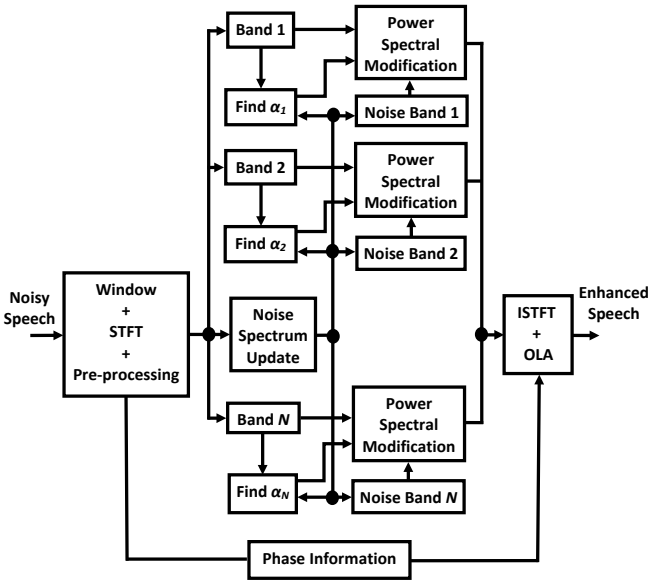

Figure 5. Block diagram of multi-band spectral subtraction algorithm [31].

response lies in the range of $0 \le H(\omega) \le 1$. This filter acts as a SNR dependent attenuator.

The Wiener filter (WF) is derived from the SSF and is based on the minimum mean squared error (MMSE) between clean speech and the estimated speech. Here, it is assumed that the speech and the noise obey normal distribution and do not correlate. The implementation of a WF requires the power spectrum of the signal and the noise. However, SSF can be used as a substitute for the WF when the signal spectrum is not available. The gain of the WF [4, 20], $H_{\text{wiener}}(\omega)$, can be expressed in terms of the power spectrum of clean speech $P_s(\omega)$ and the power spectrum of noise $P_d(\omega)$. But power spectrum of clean speech is not known, the power spectrum of the noisy speech signal $P_y(\omega)$ is used instead as

$$H_{\text{wiener}}(\omega) = \frac{P_s(\omega)}{P_y(\omega)} = \frac{P_s(\omega)}{P_s(\omega) + P_d(\omega)} = \frac{P_y(\omega) - P_d(\omega)}{P_y(\omega)} \quad (29)$$

The weakness of the WF is that it has fixed frequency response at all frequencies and the requirement to estimate the power spectral density of the clean signal and noise prior to filtering. Therefore, non-causal WF cannot be applied directly to estimate the clean speech since speech cannot be assumed to be stationary. Therefore, an adaptive WF implementation can be used to approximate (29) as

$$H_{\text{A.wiener}}(\omega) = \frac{|Y(\omega)|^2 - |\hat{D}(\omega)|^2}{|\hat{D}(\omega)|^2} \tag{30}$$

$$|\hat{S}(\omega)|^2 = H_{\text{A.wiener}}(\omega).|Y(\omega)|^2 \tag{31}$$

On comparing $H(\omega)$ and $H_{\text{A.wiener}}(\omega)$ from (28) and (30), it can be observed that the WF is based on the ensemble average spectra of the signal and noise, whereas the SSF ($b = 2$) uses the instantaneous spectra for noise signal and the time-averaged spectra of the noise. In WF theory the averaging operations are taken across the ensemble of different realization of the signal and noise processes. Whereas, in spectral subtraction we have access only to single realization of the process.

G. Iterative Spectral Subtraction Algorithm

An iterative spectral subtraction algorithm [21, 32-34] is motivated from WF [4, 20]. In this technique, the output of the spectral subtraction method is used as the input signal of the next iteration process. As after the spectral subtraction process, the type of the additive noise is changed to the remnant noise. This remnant noise is re-estimated and this new estimated noise furthermore, is been used to process the next spectral subtraction. Therefore, an enhanced output speech signal can be obtained, and the iteration process goes on. If we regard the process of noise estimate and the spectral subtraction as a filter, the filtered output is used not only for designing the filter but also as the input of the next iteration process.

The iteration time is the most important factor of this method which effects on the performance of speech enhancement. The larger iteration number will correspond to the better speech enhancement performance with the less remnant noise [31, 33].

H. Extended Spectral Subtraction Algorithm

The extended spectral subtraction [22] is based on a combination of adaptive WF and spectral subtraction, and removes the necessity of VAD to estimates the average noise spectrum during speech pauses frames. The key feature of this technique is that it can estimate average noise spectrum continuously even during speech activity without finding speech pause. WF is used to estimate the average noise spectrum and the enhanced speech spectrum is obtained by subtracting the preceding average noise spectrum from the noisy speech spectrum. This algorithm is comparatively much simpler as compared to the other subtractive-type algorithms.

I. Spectral Subtraction based on Perceptual Properties

The main weakness of spectral over-subtraction algorithm is that it uses the fixed values of subtraction parameters [15]. However, the optimization of the parameters is not an easy task, because the spectrum of most of the additive noise is not flat. An example of adaptation is multi-band spectral subtraction and parametric spectral subtraction algorithms, these schemes adapt the subtractive parameters α and β in time and frequency based on the segmental SNR or in a MSE sense, leading to improved results but remnant noise is not suppressed completely, at low SNR's [16, 19]. For this reason, the selection of proper value of parameters α and β is the major task in subtractive-type algorithms.

The concept of masking threshold of human auditory system is explored in [23], to reduce the annoying remnant noise below the noise masking threshold of lean speech signal and to make less speech distortion. In this approach, the subtraction parameters are adapted based on the noise masking threshold of human auditory system to achieve a good trade-off between the remnant noise, speech distortion and background noise. If the masking threshold is high, the remnant noise will be masked naturally and it will not be audible. In this case, the subtraction parameters have their minimum values, thereby reducing speech distortion. However, if the masking threshold is low, the remnant noise is not masked. In this case, it is necessary to increase the values of subtractive parameters. The adaptation of subtractive parameters is done according to the relations as

$$\alpha = \begin{cases} \alpha_{\max}, & \text{if } T(\omega) = T(\omega)_{\min} \\ \alpha_{\min}, & \text{if } T(\omega) = T(\omega)_{\max} \\ \alpha_{\max}\left(\dfrac{T(\omega)_{\max} - T(\omega)}{T(\omega)_{\max} - T(\omega)_{\min}}\right) + \alpha_{min}\left(\dfrac{T(\omega) - T(\omega)_{\min}}{T(\omega)_{\max} - T(\omega)_{\min}}\right), & \\ \text{if } T(\omega)\epsilon[T(\omega)_{\min}, T(\omega)_{\max}] & \end{cases} \tag{32}$$

$$\beta = \begin{cases} \beta_{\max}, & \text{if } T(\omega) = T(\omega)_{\min} \\ \beta_{\min}, & \text{if } T(\omega) = T(\omega)_{\max} \\ \beta_{\max}\left(\dfrac{T(\omega)_{\max} - T(\omega)}{T(\omega)_{\max} - T(\omega)_{\min}}\right) + \beta_{\min}\left(\dfrac{T(\omega) - T(\omega)_{\min}}{T(\omega)_{\max} - T(\omega)_{\min}}\right), & \\ \text{if } T(\omega)\epsilon[T(\omega)_{\min}, T(\omega)_{\max}] & \end{cases} \tag{33}$$

where $\alpha_{\max}, \alpha_{\min}, \beta_{\max}, \beta_{\min}$ and $T(\omega)_{\max}, T(\omega)_{\min}$ are the maximal and minimal values of α, β and updated masking threshold $T(\omega)$, respectively. It can be seen from (32) and (33) that α, β achieves the maximal and the minimal values when $T(\omega)$ equals its minimal and maximal values. The noise masking threshold can be calculated from the enhanced speech as the method proposed by [35]. The perceptual properties of human auditory system have been applied in different configuration in spectral domain and wavelet domain [36-37]. In Figure 6 the block diagram of spectral subtraction algorithm based on perceptual properties is shown.

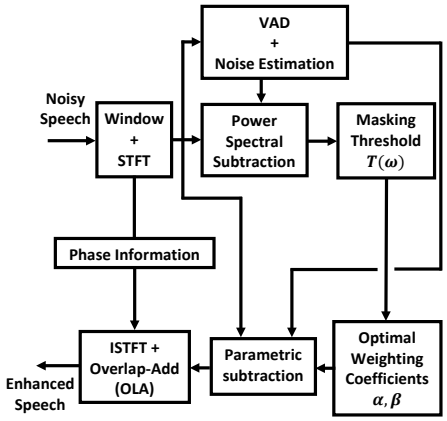

Figure 6. Block diagram of spectral subtraction algorithm based on perceptual properties.

IV. CONCLUSION

In this paper, an attempt has been made to present the comprehensive overview of research on speech enhancement using spectral subtractive-type algorithms and to provide a year-wise progress to this date. Although significant progress has been achieved in the last few decades, there are many problems yet to be resolved for enhancement of noisy speech.

Three factors that deeply affect the performance of subtractive-type algorithms are: i) exact noise estimation, ii) processing of negative spectral components, and iii) the power exponent factor. Among them the basic problem of subtractive-type algorithms are the average noise estimation. Subtraction of large amount of estimated noise makes the enhanced speech distorted, whereas subtraction of fewer amounts of estimated noise then much of the interfering noise remains present in the signal. Although the modified forms of spectral subtraction method suppress the remnant musical noise to some extent, its complete removal has not yet been achieved.

Moreover, the spectral subtractive-type algorithms are capable of enhancing the quality of speech signal but fail to enhance the intelligibility of the signal. Although, the researchers primarily focused on reducing the background noise from the degraded speech till now, yet there is a tremendous scope in enhancing the speech intelligibility and boosting up of the speech components.

REFERENCES

[1]. Y. Ephraim, H. L. Ari, and W. Roberts, "A brief survey of speech enhancement," in *the Electrical Engineering Handbook*, 3rd ed. Boca Raton, FL: CRC, 2006.

[2]. Y. Ephraim, and I. Cohen, "Recent advancements in speech enhancement," in *the Electrical Engineering Handbook*, CRC press, ch. 5, pp. 12 – 26, 2006.

[3]. Y. Ephraim, "Statistical-model-based speech enhancement systems," in *Proceedings of the IEEE*, vol. 80, no. 10, pp. 1526 – 1555, Oct. 1992.

[4]. Yifan Gong, "Speech recognition in noisy environments: A survey," *Speech Communication*, vol. 16, no. 3, pp. 261 – 291, April 1995.

[5]. J. S. Lim, and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," in *Proceedings of the IEEE*, Dec. 1979, vol. 67, no. 12, pp. 1586 – 1604.

[6]. L. W. David, and J. S. Lim, "The unimportance of phase in speech enhancement," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 30, no. 4, pp. 679 – 681, Aug. 1982.

[7]. S. F. Boll, "Suppression of noise in speech using the saber method," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, April 1978, vol. 3, pp. 606 – 609.

[8]. Y. Ephraim, and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109 – 1121, Dec. 1984.

[9]. S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113 – 120, 1979.

[10]. S. F. Boll, "A spectral subtraction algorithm for suppression of acoustic noise in speech," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, April 1979, vol. 4, pp. 200 – 203.

[11]. P. C. Loizou, *Speech Enhancement: Theory and Practice*, Boca Raton, FL: CRC, 2007.

[12]. Kuldip Paliwal, Kamil Wo´jcicki, and Belinda Schwerin, "Single channel speech enhancement using spectral subtraction in the short-time modulation domain," *Speech Communication*, vol. 52, no. 5, pp. 450 – 475, May 2010.

[13]. S. V. Vaseghi, *Advanced Digital Signal Processing and Noise Reduction*, IInd ed. NY, USA:Wiley, 2000.

[14]. Leigh D. Alsteris, and Kuldip K. Paliwal, "Short-time phase spectrum in speech processing: A review and some experimental results," *Digital Signal Processing*, vol. 17, no. 3, pp. 578 – 616, May 2007.

[15]. M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, Washington DC, April 1979, vol. 4, pp. 208 – 211.

[16]. B. L. Sim, Y. C. Tong, J. S. Chang, and C. T. Tan, "A parametric formulation of the generalized spectral subtraction method," *IEEE Transactions on Speech, and Audio Processing*, vol. 6, no.4, pp. 328 – 337, July 1998.

[17]. Yi. Hu, M. Bhatnagar, and P. C. Loizou, "A cross-correlation technique for enhancing speech corrupted with correlated noise," in *Proceedings of*

International Conference on Acoustics, Speech, and Signal Processing, May 2001, vol. 1, pp. 673 – 676.

[18]. P. Lockwood, and J. Boudy, "Experiments with a non-linear spectral subtractor (NSS), Hidden Markov Models and the projection, for robust speech recognition in cars," *Speech Communication* , vol. 11, no. 2-3, pp. 215 – 228, 1992.

[19]. S. Kamath, and P. C. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, Orlando, USA, May 2002, vol. 4, pp. 4160 – 4164.

[20]. M. A. Abd El-Fattah, M. I. Dessouky, S. M. Diaband F. E. Abd El-samie, "Speech enhancement using an adaptive wiener filtering approach," *Progress In Electromagnetics Research M.*, vol. 4, pp. 167 – 184, 2008.

[21]. S. Ogata, and T. Shimamura, "Reinforced spectral subtraction method to enhance speech signal," in *Proceedings of International Conference on Electrical and Electronic Technology*, 2001, vol. 1, pp. 242 – 245.

[22]. P. Sovka, P. Pollak, and J. Kibic, "Extended spectral subtraction," in *Proceedings of European Conference on Speech Process Communication*, Sept. 1996, pp. 963 – 966.

[23]. N. Virag, "Single-channel speech enhancement based on masking properties of the human auditory system," *IEEE Transactions on Speech, and Audio Processing*, vol. 7, pp. 126 – 137, March 1999.

[24]. R. Martin, "Spectral subtraction based on minimum statistics," in *Proceedings of European Conference on Signal Processing*, U.K., Sept. 1994, pp. 1182 – 1185.

[25]. 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, pp. 504 – 512, 2001.

[26]. R. M. Uderea, 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, pp. 1299 – 1303, 2008.

[27]. Sheng Li, Jian Qi Wang, and Xi Jing Jing, "The application of non-linear spectral subtraction method on millimeter wave conducted speech enhancement," *Mathematical Problems in Engineering*, pp. 1 – 12, 2010.

[28]. H. Tasmaz, and E. Ercelebi, "Speech enhancement based on un-decimated wavelet packet perceptual filterbanks and MMSE-STSA estimation in various noise environments," *Digital Signal processing*, vol. 18, no. 5, pp. 797 – 812, Sept. 2008.

[29]. Chao Li, and Wen-Ju Liu, "A novel multi-band spectral subtraction method based on phase modification and magnitude compensation," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, Prague, Czech Republic, May 22 – 27, 2011, pp. 4760-4763.

[30]. L. Singh, and S. Sridharan, "Speech enhancement using critical band spectral subtraction," in *Proceedings of International Conference on Spoken Language Processing*, Sydney, Australia, Dec. 1998, pp. 2827 – 2830.

[31]. Y. Ghanbari, M. R. Karmi-Mollaei, and B. Amelifard, "Improved multi-band spectral subtraction method for speech enhancement," in *Proceedings of International Conference of Signal, and Image Processing*, Hawaii, USA, Aug. 23 - 25, 2004.

[32]. K. Yamashita, S. Ogata, and T. Shimamura, "Improved spectral subtraction utilizing iterative processing," *Electronics and Communications*, Japan, vol. 90, no. 4, pp. 39 – 51, 2007.

[33]. K. Yamashita, S. Ogata, and T. Shimamura, "Spectral subtraction iterated with weighting factors," in *Proceedings of IEEE Speech Coding Workshop*, Oct. 6 - 9, 2002, pp.138 – 140.

[34]. Sheng Li, Jian-Qi Wang, Ming Niu, Xi-Jing Jing, and Tian Liu, "Iterative spectral subtraction method for millimeter wave conducted speech enhancement," *Journal of Biomedical Science and Engineering*, vol. 3, no. 2, pp. 187 – 192, Feb. 2010.

[35]. J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE Journal on Selected Areas of Communications*, vol. 6, no. 2, pp. 314 – 323, Feb. 1988.

[36]. R. M. Uderea, N. D. Vizireanu, and S. Ciochina, "An improved spectral subtraction method for speech enhancement using a perceptual weighting filter," *Digital Signal Processing*, vol. 18, pp. 581 – 587, 2008.

[37]. J. Lim, "Evaluation of a correlation subtraction method for enhancing speech degraded by additive noise," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 6, pp. 471 – 472, 1978.

Navneet Upadhyay, received the B.E. degree in electronics & communication engineering discipline from Dr. B. R. Ambedkar University, Agra, India, in 2000, and the M.Tech degree in digital communication discipline from Uttar Pradesh Technical University, Lucknow, India in 2006. Currently, he is pursuing his doctoral studies in speech signal processing engineering department of Birla Institute of Technology & Science, Pilani, India.

His research interests are in the areas of speech processing (particularly speech enhancement, speech recognition, and speech coding) and digital communication.

Abhijit Karmakar was born in West Bengal, India, in 1971. He received the B.E. degree in electronics & telecommunication engineering from Jadavpur University, India, in 1993, the M.Tech degree in electrical engineering from Indian Institute of Technology Madras, Chennai, India, in 1995, and the Ph.D degree in electrical engineering from Indian Institute of Technology, Delhi, India, in 2007.

Since 1995, he is associated with Central Electronics Engineering Research Institute/Council of Scientific & Industrial Research, Pilani, India. His research interests are is the areas of digital signal processing, auditory modeling, speech quality evaluation, and VLSI design.

