Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 54 (2015) 574 – 584

Procedia Computer Science

Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015)

# Speech Enhancement using Spectral Subtraction-type Algorithms: A Comparison and Simulation Study

Navneet Upadhyay[a,*] and Abhijit Karmakar[b]

[a] *Department of Electronics & Communication Engineering, The LNM Institute of Information Technology, Jaipur 302 031, India*
[b] *Integrated Circuit Design Group, CSIR, Central Electronics Engineering Research Institute, Pilani 333 031, India*

**Abstract**

The spectral subtraction is historically one of the first algorithms proposed for the enhancement of single channel speech. In this method, the noise spectrum is estimated during speech pauses, and is subtracted from the noisy speech spectrum to estimate the clean speech. 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. The drawback of this method is the presence of processing distortions, called remnant noise. A number of variations of the method have been developed over the past years to address the drawback. These variants form a family of spectral subtractive-type algorithms. The aim of this paper is to provide a comparison and simulation study of the different forms of subtraction-type algorithms *viz.* basic spectral subtraction, spectral over-subtraction, multi-band spectral subtraction, Wiener filtering, iterative spectral subtraction, and spectral subtraction based on perceptual properties. To test the performance of the subtractive-type algorithms, the objective measures (SNR and PESQ), spectrograms and informal listening tests are conducted for both stationary and non-stationary noises types at different SNRs levels. It is evident from the results that the modified forms of spectral subtraction method reduces remnant noise significantly and the enhanced speech contains minimal speech distortion.


Peer-review under responsibility of organizing committee of the Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015)

*Keywords:* Speech enhancement; Noise estimation; Spectral subtractive-type algorithms; Remnant noise; Objective evaluation; Spectrograms.

## 1. Introduction

Speech communication is the exchange of information *via* speech either between humans or between human to machine in the various fields' for instance automatic speech recognition and speaker identification[1]. In many situations, speech signals are degraded by the ambient noises that limit their effectiveness of communication. Therefore enhancement of speech is normally required to reduce annoyance due to noise[2]. The main purpose of speech enhancement is to decrease the distortion of the desired speech signal and to improve one or more perceptual aspects of speech, such as the quality and/or intelligibility[3]. These two measures are not necessarily correlated. Therefore, an increase in speech quality does not necessarily lead to an improvement in intelligibility[4].

Speech enhancement techniques can be classified into, single channel, dual channel or multi-channel enhancement. Although the performance of multi-channel speech enhancement is better than that of single channel enhancement[3], the

*Corresponding author. Tel.: +91-988-713-9138.
*E-mail address:* nupadhyay@lnmiit.ac.in


Peer-review under responsibility of organizing committee of the Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015)
doi:10.1016/j.procs.2015.06.066

single channel speech enhancement is still a significant field of research interest because of its simple implementation and ease of computation. In single channel applications, only a single microphone is available and the characterization of noise statistics is extracted during the periods of pauses, which requires a stationary assumption of the background noise. The estimation of the spectral amplitude of the noise data is easier than estimation of both the amplitude and phase. In[5,6], it is revealed that the short-time spectral amplitude (STSA) is more important than the phase information for the quality and intelligibility of speech.

Based on the STSA estimation, the single channel enhancement technique can be divided into two classes. The first class attempts to estimate the short-time spectral magnitude of the speech by subtracting a noise estimate. The noise is estimated during speech pauses of the noisy speech[5,6]. The second 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[6]. These two classes belong to the family of spectral subtractive-type algorithms[7,21].

The spectral subtraction method of single channel speech enhancement is the most widely used conventional method for reducing additive noise[8]. Many improvements are proposed to deal with the problems typically associated to spectral subtraction such as remnant broadband noise and narrow band tonal noise referred as musical noise[14]. In this paper, a simulation study of different forms of spectral subtractive-type algorithms is described. Other variants of spectral subtraction include spectral over-subtraction[9], multi-band spectral subtraction[10], Wiener filtering[11], iterative spectral subtraction[12], and spectral subtraction based on perceptual properties[13].

The rest of the paper is organized as follows: in Section 2, we describe the principle of the spectral subtraction method[8]. In Section 3, different forms of spectral subtractive-type algorithms[8–13] are presented. The experimental results are presented in Section 4, followed by the conclusion in Section 5.

## 2. Principle of Spectral Subtraction Method

Consider a noisy signal which consists of the clean speech degraded by statistically independent additive noise as

$$y[n] = s[n] + d[n] \tag{1}$$

where $y[n]$, $s[n]$ and $d[n]$ are the sampled noisy speech, clean speech, and additive noise, respectively. It is assumed that additive noise is zero mean and uncorrelated with the clean speech. Because the speech signal is non-stationary and time variant, the noisy speech signal is often processed on a frame-by-frame. Their representation in the short-time Fourier transform (STFT) domain is given by

$$Y(\omega, k) = S(\omega, k) + D(\omega, k) \tag{2}$$

where $k$ is a frame number. Throughout this paper, it is assumed that the speech signal is segmented into frames, hence for simplicity, we drop $k$.

Since the speech is assumed to be uncorrelated with the background noise, the short-term power spectrum of $y[n]$ has no cross-terms. Hence,

$$|Y(\omega)|^2 = |S(\omega)|^2 + |D(\omega)|^2 \tag{3}$$

The speech can be estimated by subtracting a noise estimate from the received signal.

$$|\widehat{S}(\omega)|^2 = |Y(\omega)|^2 - |\widehat{D}(\omega)|^2 \tag{4}$$

The estimation of the noise spectrum $|\widehat{D}(\omega)|^2$ is obtained by averaging recent speech pauses frames:

$$|\widehat{D}(\omega)|^2 = \frac{1}{M} \sum_{j=0}^{M-1} |Y_{SP_j}(\omega)|^2 \tag{5}$$

where $M$ is the number of consecutive frames of speech pauses (SP). If the background noise is stationary, (5) converges to the optimal noise power spectrum estimate as a longer average is taken[8].

The spectral subtraction can also be looked at as a filter, by manipulating (4) such that it can be expressed as the product of the noisy speech spectrum and the spectral subtraction filter (SSF) as:

$$|\widehat{S}(\omega)|^2 = \left(1 - \frac{|\widehat{D}(\omega)|^2}{|Y(\omega)|^2}\right)|Y(\omega)|^2 \tag{6}$$

$$= H^2(\omega)|Y(\omega)|^2 \tag{7}$$

where $H(\omega)$ is the gain function and known spectral subtraction filter (SSF). The $H(\omega)$ is a zero phase filter, with its magnitude response in the range of $0 \leq H(\omega) \leq 1$.

$$H(\omega) = \left\{\max\left(0, 1 - \frac{|\widehat{D}(\omega)|^2}{|Y(\omega)|^2}\right)\right\}^{1/2} \tag{8}$$

To reconstruct the resulting signal, the phase estimate of the speech is also needed. A common phase estimation method is to adopt the phase of the noisy signal as the phase of the estimated clean speech signal, based on the notion that short-term phase is relatively unimportant to human ears[5]. Then, the speech signal in a frame is estimated as

$$\widehat{S}(\omega) = |\widehat{S}(\omega)|e^{j<Y(\omega)} = H(\omega)Y(\omega) \tag{9}$$

The estimated speech waveform is recovered in the time domain by inverse Fourier transforming $\widehat{S}(\omega)$ using an overlap and add approach[8].

The spectral subtraction method, although reducing the noise significantly, it has some severe drawbacks. From (4), it is clear that the effectiveness of spectral subtraction is heavily dependent on accurate noise estimation, which is a difficult task to achieve in most conditions. When the noise estimate is less than perfect, two major problems occur, remnant noise with musical structure and speech distortion.

## 3. Spectral Subtractive-type Algorithms

The spectral subtractive-type algorithm is the family of different variants of the spectral subtraction method such as spectral over-subtraction, multi-band spectral subtraction, Wiener filtering, iterative spectral subtraction, and spectral subtraction based on perceptual properties. Thus, the principle of the spectral subtractive-type algorithms is to estimate the short-time spectral magnitude of the speech by subtracting estimated noise from the noisy speech spectrum or by multiplying the noisy spectrum with gain functions and to combine it with the phase of the noisy speech.

### *A. Spectral over-subtraction*

In this algorithm[9], two additional parameters are introduced in the spectral subtraction method[8]: over-subtraction factor, and noise spectral floor to reduce the remnant noise. The algorithm is given as

$$|\widehat{S}(\omega)|^2 = \begin{cases} |Y(\omega)|^2 - \alpha|\widehat{D}(\omega)|^2, & \text{if } |Y(\omega)|^2 > (\alpha+\beta)|\widehat{D}(\omega)|^2 \\ \beta|\widehat{D}(\omega)|^2 & \text{else} \end{cases} \tag{10}$$

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 prevent the resultant spectrum from going below a preset minimum level rather than setting to zero (spectral floor). The over-subtraction factor depends on a-*posteriori* segmental SNR (SSNR). The over-subtraction factor can be calculated as

$$\alpha = 4 - \frac{3}{20}\,\text{SSNR}, \quad \text{if } -5 \leq \text{SSNR} \leq 20 \tag{11}$$

$$\mathrm{SSNR} = \left( \frac{\sum_{k=0}^{NF-1} |Y(\omega)|^2}{\sum_{k=0}^{NF-1} |\widehat{D}(\omega)|^2} \right) \tag{12}$$

Here $NF$ is the number of frames in the signal.

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 background noise and remnant noise removal, various combinations of over-subtraction factor $\alpha$, and spectral floor parameter $\beta$ give rise to a trade-off between the amount of remaining background noise and the level of perceived remnant noise. For large values of $\beta$, the spectral floor is high, and a very little, if any remnant noise is audible, while with small $\beta$, the background noise is greatly reduced, but the remnant noise becomes quite annoying. Hence, the suitable value of $\alpha$ is set as (11) and $\beta = 0.03$.

This algorithm reduces the noise to some extent but the remnant noise is not completely eliminated, effecting the quality of the speech signal. Also, the algorithm assumes that the noise affects the whole speech spectrum equally. Consequently, it uses a single value of the over-subtraction factor for the whole speech spectrum. Therefore, the enhanced speech is distorted.

*B. Multi-band spectral subtraction*

Real world noise is mostly colored and affects the speech signal differently over the entire spectrum. This is illustrated in Figure 1, which is the plot of SSNR of non-overlapped uniformly spaced frequency bands {60 Hz $\sim$ 1 kHz (Band 1), 1 kHz $\sim$ 2 kHz (Band 2), 2 kHz $\sim$ 3 kHz (Band3), 3 kHz $\sim$ 4kHz (Band 4)} over frame number. This figure shows that the SSNR of the low frequency bands (Band 1) is significantly higher than the SSNR of higher frequency bands (Band 4)[6,10,16]. Therefore, the use of frequency dependent subtraction factor to account for different types of noise. The idea of non-linear spectral subtraction (NSS)[7], basically extend this capability by making the over-subtraction factor frequency dependent and subtraction process is non-linear. Larger values are subtracted at frequencies with low SNR levels, and smaller values are subtracted at frequencies with high SNR levels. Certainly, this gives higher flexibility in compensating for errors in estimating the noise energy in different frequency bins.

To take into account, a uniformly frequency spaced multi-band approach to spectral subtraction was presented in[10]. In this algorithm, the speech spectrum is divided into four uniformly spaced frequency bands, and spectral subtraction is performed independently in each band. The algorithm re-adjusts the over-subtraction factor in each band based on SSNR. So, the estimate of the clean speech magnitude spectrum in the $i^{\text{th}}$ Band is obtained by:

$$|\widehat{S}_i(\omega)|^2 = \begin{cases} |Y_i(\omega)|^2 - \alpha_i \, \delta_i |\widehat{D}_i(\omega)|^2, & \text{if } |\widehat{S}_i(\omega)|^2 > 0 \; k_i < \omega < k_{i+1} \\ \beta |Y_i(\omega)|^2 & \text{else} \end{cases} \tag{13}$$

where $k_i$ and $k_{i+1}$ are the start and end frequency bins of the $i^{\text{th}}$ frequency band, $\alpha_i$ is the band specific over-subtraction factor of the $i^{\text{th}}$ Band, which is the function of SSNR of the $i^{\text{th}}$ frequency band. The SSNR of the $i^{\text{th}}$ frequency band can be calculated as

$$\mathrm{SSNR}_i(\omega) = \left( \frac{\sum_{\omega=k_i}^{k_{i+1}} |Y_i(\omega)|^2}{\sum_{\omega=k_i}^{k_{i+1}} |\widehat{D}_i(\omega)|^2} \right) \tag{14}$$

The band specific over-subtraction can be calculated, as

$$\alpha_i = \begin{cases} 5, & \text{if } \mathrm{SNR}_i \leq -5 \\ 4 - \frac{3}{20} \, \mathrm{SSNR}_i, & \text{if } -5 \leq \mathrm{SNR}_i \leq 20 \\ 1, & \text{if } \mathrm{SNR}_i > 20 \end{cases} \tag{15}$$

The $\delta_i$ 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.

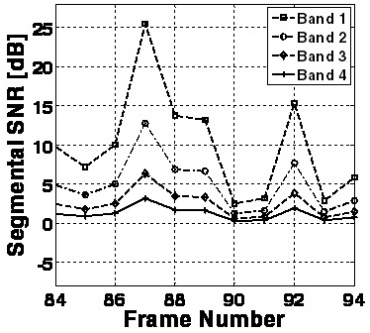


Fig. 1. The segmental SNR of bands[7,21].

The values of $\delta_i$[10] is empirically calculated and set to

$$\delta_i = \begin{cases} 1, & f_i \leq 1\,\text{kHz} \\ 2.5, & 1\,\text{kHz} < f_i \leq \frac{f_s}{2} - 2\,\text{kHz} \\ 1.5, & f_i > \frac{f_s}{2} - 2\,\text{kHz} \end{cases} \tag{16}$$

Here $f_i$ is the upper bound frequency of the $i^{\text{th}}$ Band and $f_s$ is the sampling frequency. The motivation for using smaller values of $\delta_i$ for the low frequency bands is to minimize speech distortion, since most of the speech energy is present in the lower frequencies. Both factors, $alpha_i$ and $\delta_i$ 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. The MBSS algorithm is found to perform better than other subtractive-type algorithms[8–10].

*C. Wiener filtering*

The Wiener filter (WF) is an optimal filter that minimizes the mean square error criterion[5,11]. Here, it is assumed that the speech and the noise obey normal distribution and do not correlate. The gain function of WF, $H_{\text{wiener}}(\omega)$, can be expressed in terms of the power spectral density of clean speech $P_s(\omega)$ and the power spectral density of noise $P_d(\omega)$[5,11] as

$$H_{\text{wiener}}(\omega) = \frac{P_s(\omega)}{P_s(\omega) + P_d(\omega)} \tag{17}$$

The weakness of the WF is that the fixed gain function 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 (17) as

$$H_{\text{A. wiener}}(\omega) = \frac{|\widehat{S}(\omega)|^2}{|Y(\omega)|^2} \tag{18}$$

$$|\widehat{S}(\omega)|^2 = H_{\text{A.wiener}}(\omega)|Y(\omega)|^2 \tag{19}$$

$H_{\text{A.wiener}}(\omega)$ attenuates each frequency component by a certain amount depending on the power of the noise at the frequency.

If $|\widehat{D}(\omega)|^2 = 0$, then $H_{\text{A.wiener}}(\omega) = 1$ and no attenuation takes place, whereas if $|\widehat{D}(\omega)|^2 = |Y(\omega)|^2$, then $H_{\text{A.wiener}}(\omega) = 0$. Therefore, the frequency component is completely nulled. All other values of $H_{\text{A.wiener}}(\omega)$ scale the power of the signal by an appropriate amount.

On comparing $H(\omega)$ and $H_{\text{A.wiener}}(\omega)$ from (8) and (18), it can be observed that the WF is based on the ensemble average spectra of the signal and noise, whereas the SSF uses the instantaneous spectra for noise signal and the running average (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. In spectral subtraction, we have access only to the single realization of the process.

Using of power spectrum of noisy speech, instead of that of clean speech for calculating the gain function degrades WF accuracy. To solve this problem, an iterative algorithm is used[5].

*D. Iterative spectral subtraction*

An iterative spectral subtraction (ISS) algorithm is proposed in[12] which is motivated from WF[5,11], to suppress the remnant noise. In this algorithm, the output of the enhanced speech is used as the input signal for the next iteration process. As after the spectral subtraction process, the type of the additive noise is transformed to the remnant noise and the output signal is used as the input signal of the next iteration process. The remnant noise is re-estimated and this new estimated noise, furthermore, is used to process the next spectral subtraction process. 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.

Moreover, the iteration number is the most important factor of this algorithm which affects the performance of speech enhancement system. Therefore, the larger iteration number corresponds to better speech enhancement with the less remnant noise[19,20].

*Spectral subtraction based on perceptual properties*

The main weakness of spectral subtraction[9] is that it uses the fixed value of subtraction parameters that are unable to adapt the variable noise-levels and noise characteristics. However, the optimization of the parameters is not an easy task, because the spectrum of most of the noise, added in speech, is not flat. An example of adaptation is multi-band spectral subtraction, which adapts the subtractive parameters in time and frequency based on the SSNR, leading to improved results, but remnant noise are not suppressed completely at low SNR's[10]. Therefore, the selection of the appropriate value of subtractive parameters is the major task in subtractive-type algorithms for enhancement of noisy speech.

The spectral subtraction based on perceptual properties has been investigated to improve intelligibility and quality of the speech signals[13]. The masking properties of human auditory system are incorporated into the enhancement process in order to attenuate the noise components that are already inaudible due to masking. In the algorithm[13], the subtraction parameters are adapted based on the masking properties. The masking properties are modelled by calculating the noise masking threshold[17]. A human listener tolerates additive noise as long as it remains below this threshold. The adaptation of subtraction parameters is done according to the relations

$$\alpha = \begin{cases} \alpha_{\max}, & \text{if } T(\omega) = T(\omega)_{\min} \\ \alpha_{\min}, & \text{if } T(\omega) = T(\omega)_{\max} \\ \alpha_{\max}\left(\frac{T(\omega)_{\max} - T(\omega)}{T(\omega)_{\max} - T(\omega)_{\min}}\right) + \alpha_{\min}\left(\frac{T(\omega) - T(\omega)_{\min}}{T(\omega)_{\max} - T(\omega)_{\min}}\right), & \text{if } T(\omega) \in [T(\omega)_{\min}, T(\omega)_{\max}] \end{cases} \tag{20}$$

$$\beta = \begin{cases} \beta_{\max}, & \text{if } T(\omega) = T(\omega)_{\min} \\ \beta_{\min}, & \text{if } T(\omega) = T(\omega)_{\max} \\ \beta_{\max}\left(\frac{T(\omega)_{\max} - T(\omega)}{T(\omega)_{\max} - T(\omega)_{\min}}\right) + \beta_{\min}\left(\frac{T(\omega) - T(\omega)_{\min}}{T(\omega)_{\max} - T(\omega)_{\min}}\right), & \text{if } T(\omega) \in [T(\omega)_{\min}, T(\omega)_{\max}] \end{cases} \tag{21}$$

Here $\alpha_{\max}, \alpha_{\min}, \beta_{\max}, \beta_{\min}$ and $T(\omega)_{\max}, T(\omega)_{\min}$ are the maximal and minimal values of $\alpha, \beta$ and updated masking threshold $T(\omega)$ respectively[13]. It can be seen from (21) and (22) that $\alpha, \beta$ achieves the maximal and the minimal

values when $T(\omega)$ equalize its minimal and maximal values. The noise masking threshold can be calculated from the enhanced speech as the method proposed by[17].

## 4. Experimental Results

In this section, the each variant of spectral subtraction method is evaluated and compared with other variants. The speech datasets used in our simulations are from the NOIZEUS corpus[18]. The NOIZEUS composed of 30 phonetically balanced sentences pronounced by six speakers (three male and three female) in English language. The corpus is sampled at 8 kHz and filtered to simulate receiving frequency characteristics of telephone handsets. Noise signals have different time-frequency distributions, and therefore a different impact on clean signal. For that reason, the NOIZEUS comes with various non-stationary noises at different levels of SNRs. The non-stationary noises are car, train, restaurant, babble, airport, street, and exhibition.

In our evaluation, we have used the speech degraded by car noise at global SNR levels of 0 dB to 15 dB in steps of 5 dB. We also generate a corresponding stimulus set degraded by additive white Gaussian noise (AWGN), stationary noise, at four SNR levels: 0 dB, 5 dB, 10 dB, and 15 dB. The performance of the subtractive-type algorithms, tests on such noisy speech samples.

In our experiments, the noise samples used are of zero-mean and the energy of the noisy speech samples are normalized to unity. The frame size is chosen to be 256 samples with 50% overlap. The sinusoidal Hamming window with size 256 samples is applied to each frame before it is enhanced individually. The noise estimate is updated during the silence frames by using averaging. The final enhanced speech is reconstructed from the enhanced frames using the weighted overlap and adds technique.

For SOS algorithm, the value of $\alpha$ is set as (11) and $\beta$ is kept fixed at 0.03. For MBSS approach, four linearly frequencies spaced bands is used with $\beta = 0.03$ and the value of $\alpha_i$ and $\delta_i$ is set as (15) and (16). For WF, the value of smoothing constant is taken as 0.99. For ISS algorithm, the iteration time is taken as 2–3 and for SSPP algorithm the value of $\alpha_{\max} = 6, \alpha_{\min} = 1$, $\beta_{\min} = 0$, and $\beta_{\max} = 0.02$[15].

The SNR improvement is the performance evaluation for calculating the amount of noise reduction in the background noise level conditions. The obtained value of SNR improvement for WGN of different enhancement algorithms is presented in Fig. 2. The better noise reduction and least speech distortion is obtained in case of SSPP algorithm compared to other algorithms. The main drawback of the SNR is the fact that it has a poor correlation with subjective quality assessment results. Therefore, the SNR of enhanced speech is not a sufficient objective indicator of speech quality.

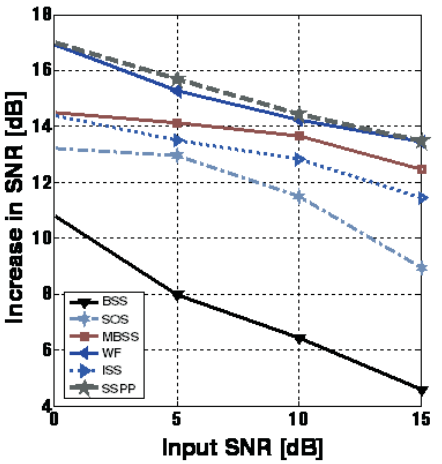


Fig. 2. The improved SNR of different subtractive-type algorithms for WGN.

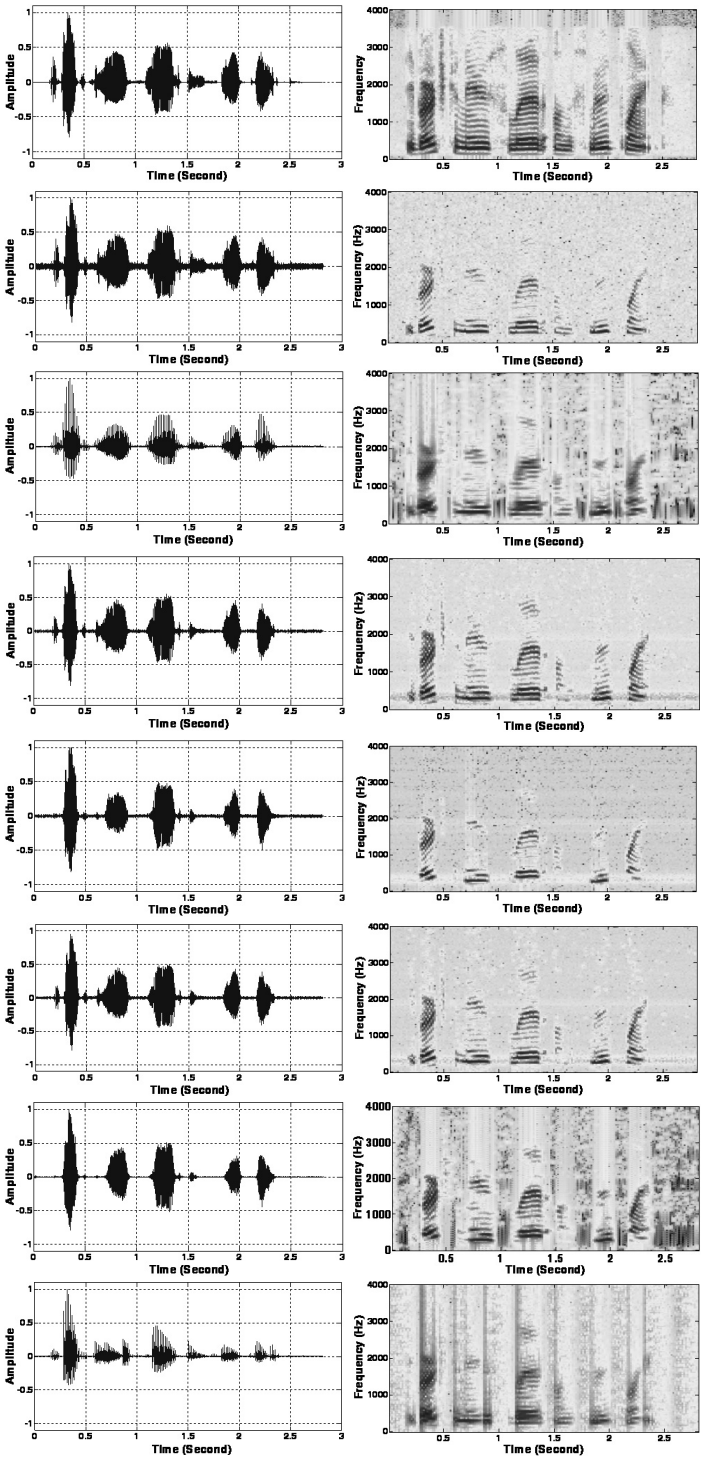


Fig. 3. Waveforms and spectrograms (From top to bottom): (i) Clean speech; (ii) Noisy speech (white noise at 15 dB); (iii)–(viii) Speech enhanced by different subtractive-type algorithms; (iii) BSS (PESQ = 2.151); (iv) SOS (PESQ = 2.800); (v) MBBS (PESQ = 2.563); (vi) ISS (PESQ = 2.840; (vii) WF (PESQ = 2.910); and (viii) SSPP ((PESQ = 2.980).

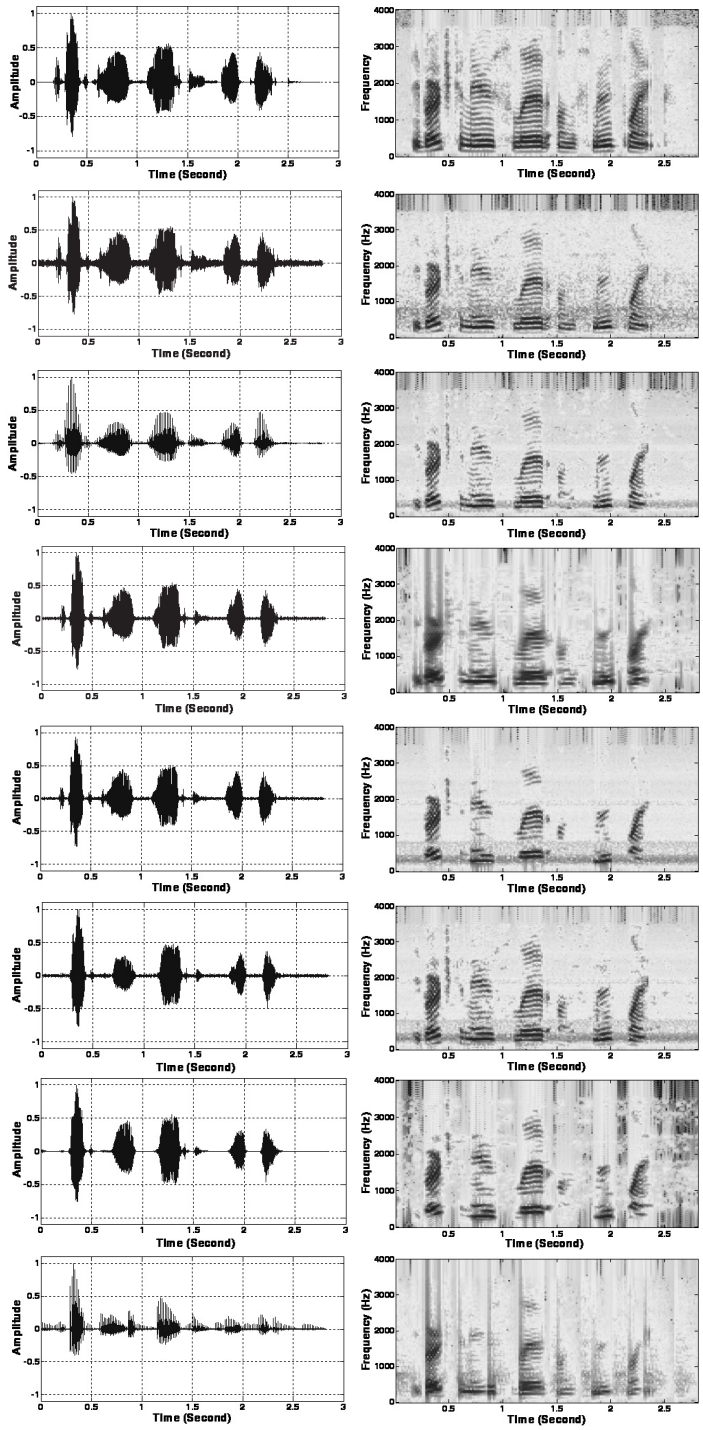


Fig. 4. Waveforms and spectrograms (From top to bottom): (i) Clean speech, (ii) Degraded speech (car noise at 15 dB); (iii)–(viii) Speech enhanced by different subtractive-type algorithms; (iii) BSS ((PESQ = 2.213), (iv) SOS (PESQ = 2.831), (v) MBBS (PESQ = 2.602), (vi) ISS (PESQ = 2.850), (vii) WF (PESQ = 2.970); and (viii) SSPP (PESQ = 3.100).

The perceptual evaluation of speech quality (PESQ) is an objective quality measure designed to predict the subjective opinion score of a degraded audio sample and it is recommended by ITU-T for speech quality assessment[22]. 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[22] for a large number of testing conditions. The PESQ is one of the best measures of signal's quality. The PESQ score of enhanced speech by subtractive-type algorithms is shown in Fig. 3 and Fig. 4.

Normally, spectral subtractive-type speech enhancement algorithms generate two main undesirable effects: remnant noise and speech distortion. These two effects can be annoying to a human listener, and causes listener fatigue. However, they are difficult to quantify. Therefore, it is important to analyze the time-frequency distribution of the enhanced speech, in particular the musical structure of its remnant noise. The speech spectrogram is a good tool to do this work, because it can give more accurate information about remnant noise and speech distortion than the corresponding time domain waveforms. For comparison purpose, Figure 3 shows the plot of temporal waveforms and spectrograms of the clean speech signal, noisy speech (degraded by WGN) and speech enhanced by the different spectral subtractive-type algorithms, namely, BSS, SOS, MBSS, WF, ISS, and SSPP with PESQ score. Figure 4 shows the temporal waveforms and spectrograms of enhanced speech in case of car noise with PESQ scores.

Figure 3 (iii) presents the enhanced speech obtained by basic spectral subtraction with no remnant noise reduction. The remnant noise level is very important and its musical structure can be observed. This shows that this basic method cannot be used at very low SNR without any improvement.

Figure 3 (iv)–(viii) shows an enhanced speech spectrogram obtained with algorithms SOS, MBSS, ISS, WF, and SSPP algorithm. From the spectrograms, we can easily observe that the MBSS, ISS, and Wiener filtering have a very small amount of remnant noise and spectral subtraction based on perceptual properties has a better performance compared to other algorithms for speech enhancement. Wiener filtering results in a smaller amount of remnant noise, but this noise has musical structure and speech regions, especially fricative consonants, are also attenuated. This type of spectral subtraction can result in speech distortion. Also, in case of car noise, Fig. 4, the BSS, SOS, ISS, and WF results are weak compared to MBSS and SSPP. This is also be justified by the PESQ score of different speech enhancement algorithms.

The best results were obtained with spectral subtraction with perceptual properties. In case of this type of subtractive-type algorithm small amount of remnant noise is remaining, but this noise has a perceptually white quality and distortion remains acceptable. Informal listening tests also indicated that the enhanced speech with SSPP algorithm is more pleasant, the remnant noise is better reduced, and with minimal, if any, speech distortion.

## 5. Conclusion

In this paper, a comparison and simulation study of different forms of spectral subtractive-type algorithms for suppression of additive noise is presented. In particular, algorithms based on short-time Fourier transforms are examined and the limitations of spectral subtraction method are discussed briefly.

The performance evaluation of subtractive-type algorithms is carried out using objective measures (SNRs and PESQ score), and spectrograms with informal subjective listening tests. The results shows that the classical spectral subtraction algorithm mostly results in audible remnant noise, which decreases speech intelligibility. The most progressive algorithm of speech enhancement is the spectral subtraction based on perceptual properties. This algorithm takes advantage of how people perceive the frequencies instead of just working with SNR. It results in appropriate remnant noise suppression and acceptable degree of speech distortion.

## References

[1] D. O'Shaughnessy, Speech Communications: Human and Machine, 2nd ed. Hyderabad, India: University Press (I) Pvt. Ltd., (2007).
[2] Y. Ephraim, Statistical-Model-Based Speech Enhancement Systems, *The IEEE*, vol. 80, no. 10, pp. 1526–1555, October (1992).
[3] Y. Ephraim, H. L. Ari and W. Roberts, A Brief Survey of Speech Enhancement, The Electrical Engineering Handbook, 3rded. Boca Raton, FL: CRC, (2006).
[4] Y. Ephraim and I. Cohen, Recent Advancements in Speech Enhancement, The Electrical Engineering Handbook, CRC press, ch. 5, pp. 12–26, (2006).
[5] J. S. Lim and A. V. Oppenheim, Enhancement and Bandwidth Compression of Noisy Speech, *The IEEE*, vol. 67, pp. 1586–1604, (1979).

[6] P. C. Loizou, Speech Enhancement: Theory and Practice, I$^{st}$ ed. Taylor and Francis, (2007).
[7] Navneet Upadhyay and Abhijit Karmakar, The Spectral Subtractive-Type Algorithms for Enhancing Speech in Noisy Environments, *IEEE Int. Conf. on Recent Advances in Information Technology*, ISM Dhanbad, India, March 15–17, pp. 841–847, (2012).
[8] 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).
[9] M. Berouti, R. Schwartz and J. Makhoul, Enhancement of Speech Corrupted by Acoustic Noise, *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Washington DC, pp. 208–211, April (1979).
[10] S. Kamath and P. Loizou, A Multi-Band Spectral Subtraction Method for Enhancing Speech Corrupted by Colored Noise, *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Orlando, USA, vol. 4, pp. 4160–4164, May (2002).
[11] M. A. Abd El-Fattah, M. I. Dessouky, S. M. Diab and F. E. Abd El-samie, Speech Enhancement using an Adaptive Wiener Filtering Approach, *Progress in Electromagnetic Research M.*, vol. 4, pp. 167–184, (2008).
[12] S. Ogata and T. Shimamura, Reinforced Spectral Subtraction Method to Enhance Speech Signal, *IEEE Int. Conf. on Electrical and Electronic Technology*, vol. 1, pp. 242–245, (2001).
[13] N. Virag, Single Channel Speech Enhancement Based on Masking Properties of the Human Auditory System, *IEEE Transactions on Speech, and Audio Processing*, vol. 7, no. 2, pp. 126–137, March (1999).
[14] S. V. Vaseghi, Advanced Digital Signal Processing and Noise Reduction, 2$^{nd}$ ed. NY, USA: Wiley, (2000).
[15] P. Lockwood and J. Boudy, Experiments with a Nonlinear Spectral Subtractor (NSS), Hidden Markov Models and Projection, for Robust Recognition in Cars, *Speech Communication*, vol. 11, no. 2–3, pp. 215–228, (1992).
[16] Y. Ghanbari, M. R. K. Mollaei and B. Amelifard, Improved Multi-Band Spectral Subtraction Method for Speech Enhancement, *IEEE Int. Conf. on Signal, and Image Processing*, Hawaii, USA, August (2004).
[17] 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, February (1988).
[18] A Noisy Speech Corpus for Evaluation of Speech Enhancement Algorithms. http://www.utdallas.edu/∼loizou/speech/noizeus/.
[19] K. Yamashita, S. Ogata and T. Shimamura, Improved Spectral Subtraction Utilizing Iterative Processing, *Electronics and Communications*, Japan, Part 3, vol. 90, no. 4, pp. 39–51, (2007).
[20] 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, February (2010).
[21] Navneet Upadhyay and Abhijit Karmakar, Spectral Subtractive-Type Algorithms for Enhancement of Noisy Speech: An Integrative Review International Journal Image, *Graphics and Signal Processing*, vol. 5, no. 11, pp. 13–22, September (2013).
[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., pp. 862, (2000).