Single-Channel Speech Enhancement by NWNS and EMD

Somlal Das somlal_ru@yahoo.com
Dept. of Computer Science and Engineering
University of Rajshahi
Rajshahi, Bangladeh

Md. Ekramul Hamid ekram_hamid@yahoo.com
Department of Network Engineering
King Khalid University
Abha, Kingdom of Saudi Arabia

Keikichi Hirose hirose@gavo.t.u-tokyo.ac.jp
Dept. of Information and Communication Eng.
The University of Tokyo
Tokyo, Japan

Md. Khademul Islam Molla molla@gavo.t.u-tokyo.ac.jp
Dept. of Information and Communication Eng.
The University of Tokyo
Tokyo, Japan

Abstract

This paper presents the problem of noise reduction from observed speech by means of improving quality and/or intelligibility of the speech using single-channel speech enhancement method. In this study, we propose two approaches for speech enhancement. One is based on traditional Fourier transform using the strategy of Noise Subtraction (NS) that is equivalent to Spectral Subtraction (SS) and the other is based on the Empirical Mode Decomposition (EMD) using the strategy of adaptive thresholding. First of all, the two different methods are implemented individually and observe that, both the methods are noise dependent and capable to enhance speech signal to a certain limit. Moreover, traditional NS generates unwanted residual noise as well. We implement nonlinear weight to eliminate this effect and propose Nonlinear Weighted Noise Subtraction (NWNS) method. In first stage, we estimate the noise and then calculate the Degree Of Noise (DON1) from the ratio of the estimated noise power to the observed speech power in frame basis for different input Signal-to-Noise-Ratio (SNR) of the given speech signal. The noise is not accurately estimated using Minima Value Sequence (MVS). So the noise estimation accuracy is improved by adopting DON1 into MVS. The first stage performs well for wideband stationary noises and performed well over wide range of SNRs. Most of the real world noise is narrowband non-stationary and EMD is a powerful tool for analyzing non-linear and non-stationary signals like speech. EMD decomposes any signals into a finite number of band limited signals called intrinsic mode function (IMFs). Since the IMFs having different noise and speech energy distribution, hence each IMF has a different noise and speech variance.

These variances change for different IMFs. Therefore an adaptive threshold function is used, which is changed with newly computed variances for each IMF. In the adaptive threshold function, adaptation factor is the ratio of the square root of added noise variance to the square root of estimated noise variance. It is experimentally observed that the better speech enhancement performance is achieved for optimum adaptation factor. We tested the speech enhancement performance using only EMD based adaptive thresholding method and obtained the outcome only up to a certain limit. Therefore, further enhancement from the individual one, we propose two-stage processing technique, NWNS+EMD. The first stage is used as a pre-process for noise removal to a certain level resulting first enhanced speech and placed this into second stage for further removal of remaining noise as well as musical noise to obtain final enhancement of the speech. But traditional NS in the first stage produces better output SNR up to 10 dB input SNR. Furthermore, there are musical noise and distortion presented in the enhanced speech based on spectrograms and waveforms analysis and also from informal listening test. We use white, pink and high frequency channel noises in order to show the performance of the proposed NWNS+EMD algorithm.

Keywords: speech enhancement, non linear weighted noise subtraction, degree of noise, empirical mode decomposition, adaptive thresholding.

1. INTRODUCTION

In many speech related systems like mobile communication in an adverse environment, the desired signal is not available directly; rather it is mostly contaminated with some interference sources of noise. These background noise signals degrade the quality and intelligibility of the original speech, resulting in a severe drop in the performance of the applications. The degradation of the speech signal due to the background noise is a severe problem in speech related systems and therefore should be eliminated through speech enhancement algorithms. In our previous study, we have proposed a two stage noise reduction algorithm by noise subtraction and blind source separation [1]. In that report, we recommended further research to improve the algorithm over wide ranges of SNRs as well as noise reduction performance for narrow-band noises.

Research on speech enhancement techniques started more than 40 years ago at AT&T Bell Laboratories by Schroeder as mentioned in [2]. Schroeder proposed an analog implementation of the spectral magnitude subtraction method. Then, the method was modified by Schroeder's colleagues in a published work [3]. However, more than 15 years later, the spectral subtraction method as proposed by Boll [4] is a popular speech enhancement techniques through noise reduction due to its simple underlying concept and its effectiveness in enhancing speech degraded by additive noise. The technique is based on the direct estimation of the short-term spectral magnitude. Recent studies have focused on a non-linear approach to the subtraction procedure [5-7]. In Martin [5] algorithm modifies the short time spectral magnitude of the corrupted speech signal such that the synthesized signal is perceptually as close as possible to the clean speech signal. The estimating noise is obtained as the minima values of a smoothed power estimate of the noisy signal, multiplied by a factor that compensates the bias. The algorithm eliminates the need of speech activity detector by exploiting the short time characteristics of speech signal. Martin's study compared the result with Malah [6], and found an improved SNR. However, this noise estimation is sensitive to outliers, and its variance is about twice as large as the variance of a conventional noise estimator. These approaches have been justified due to the variation of signal-to-noise ratio across the speech spectrum. Unlike white

Gaussian noise, which has a flat spectrum, the spectrum of real-world noise is not flat. Thus, the noise signal does not affect the speech signal uniformly over the whole spectrum. Some frequencies are affected more adversely than others. In high frequency channel noise (HF channel), for instance, in the low frequencies, where most of the speech energy resides, are affected more than the high frequencies. Hence it becomes imperative to estimate a suitable factor that will subtract just the necessary amount of the noise spectrum from each frequency bin (ideally), to prevent destructive subtraction of the speech while removing most of the residual noise. Then it is usually difficult to design a standard algorithm that is able to perform homogeneously across all types of noise. For that, a speech enhancement system is based on certain assumptions and constraints that are typically dependent on the application and the environment.

There are some crucial restrictions of the Fourier spectral analysis [8]: the system must be linear; and the data must be strictly periodic or stationary; otherwise the resulting spectrum will make little physical sense. From this point of view, Fourier filter methods will fail when the processes are nonlinear. The empirical mode decomposition (EMD), proposed by Huang *et.al* [9] as a new and powerful data analysis method for nonlinear and non-stationary signals, has made a new path for speech enhancement research. EMD is a data-adaptive decomposition method, which decompose data into zero mean oscillating components, named as intrinsic mode functions (IMFs). It is mentioned in [10] that most of the noise components of a noisy speech signal are centered on the first three IMFs due to their frequency characteristics. Therefore EMD can be used for effectively identifying and removing these noise components. Xiaojie et. al. [11] proposed EMD that effectively identify and remove noise components. Recently there are many speech enhancement methods [12-14] have been developed in dual-channel and single-channel modes using EMD. In [12] EMD based speech enhancement is achieved by removing those IMFs whose energies exceeded a predefined threshold value. The IMFs, which represent empirically, observed applying EMD in observed speech contaminated with white Gaussian noise generates noise model. In [13] speech enhancement based on EMD-MMSE is performed by filtering the IMFs generated from the decomposition of speech contaminated with white Gaussian noise. In [14], an optimum gain function is estimated for each IMF to suppress residual noise that may be retained after single-channel speech enhancement algorithms.

In our previous study, Hamid [1] proposed noise subtraction (NS) technique where noise is estimated using minimum value sequence (MVS) and the noise floor is updated with the help of estimated degree of noise (DON). The main drawback of this method is that we estimate DON on the basis of pitch period over the frame and the pitch period of unvoiced sections is not accurately estimated. To solve this problem, in this paper, we estimate EDON on the basis of estimated SNRs of clean and noisy speech spectrums. Then, the EDON is estimated in two stages from a function, which is previously prepared as the function of the parameter of the degree of noise [1]. We consider the valleys of the observed smoothed power spectrum of a noisy speech signal to estimate noise power. This spectrum is tuned by EDON to adjust the noise level for a particular SNR. We also perform suitable steps to minimize the residual noise problem. Now the estimated noise spectrum with a controlled non-linear factor is subtracted from the observed spectrum in time domain to obtain noise reduced speech. This paper presents a parametric formulation to estimate noise weight on the basis of EDON. The weighting factor increases with increasing SNRs, and results non-linear weighting factor with speech activity. Although Fourier transform and wavelet analysis make great contributions, they suffer from many shortcomings in case of nonlinear and nonstationary signals. For this reason, for further enhancement, EMD technique has been used for robust noisy speech analysis in this work.

Since the IMFs in EMD having different noise and speech energy distribution, hence each IMF has a different noise and speech variance. These variances change for different IMFs. Therefore an adaptive threshold function is used, which is changed with newly computed variances for each IMF. Moreover, since IMFs are generated from EMD and therefore, we call the proposed method as EMD based adaptive thresholding technique. To enhance the speech, EMD based adaptive thresholding algorithm applied into each IMFs for removing the noise embedded in the underlying

IMFs. In the adaptive threshold function, adaptation factor is the ratio of the square root of added noise variance to the square root of estimated noise variance. It is experimentally observed that the better speech enhancement performance is achieved for optimum adaptation factor. We tested the speech enhancement performance using only EMD based adaptive thresholding method and obtained the outcome only up to a certain limit. Moreover, each individual method has some performance limitations.

Therefore, further enhancement from the individual one, we propose two-stage processing technique, namely, a time domain NS or NWNS followed by an EMD based adaptive thresholding. The first stage is used as a pre-process for noise removal to a certain level resulting first enhanced speech and placed this into second stage for further removal of remaining noise as well as musical noise to obtain final enhancement of the speech. But traditional NS in the first stage produces better output SNR up to 10 dB input SNR. Furthermore, there are musical noise and distortion presented in the enhanced speech based on spectrograms and waveforms analysis and also from informal listening test. EMD based adaptive thresholding does not work well on distorted speech and not be able to recover the speech from the distorting speech when it cascaded with NS. As a result, the overall performance of enhanced speech obtained from NS+EMD based adaptive thresholding is not so good based on the objective and subjective measures. In the first stage, the performance of speech enhancement improves by introducing nonlinear weight in NS, namely NWNS, to control the noise level and improves its overall performance for wide range of input SNRs provide first enhanced speech without distortion and with minimum effect of musical noise. Moreover, the overall performance is further improved by cascading NWNS in the first stage and EMD based adaptive thresholding in the second stage. In this two-stage processing, NWNS is influenced to increase the performance of EMD based adaptive thresholding. The advantage of the method is the effective removal of noise and produces better output SNR for wide range of input SNR and also improves the speech quality with reducing residual noise.

2. NOISE ESTIMATION AND SUBTRACTION

The main component of speech noise reduction is noise estimation that is a most difficult task for a single-channel enhancement system. The noise estimate can have a major impact on the quality of the enhanced speech. That is, with a better noise estimation, a more correct SNR is obtained, resulting in the enhanced speech with low distortion. We have assumed that speech and noise are uncorrelated to each other. We further assume that signal and noise are statistically independent.

2.1 Estimating Minimum Value Sequence (MVS)

The sections of consecutive samples are used as a single frame l(320 samples) and spaced l'(100 samples) achieving an almost 62.75% overlap. The short-term representation of a signal $y(n)$ is obtained by Hamming windowing and analyzed using N=512 point Discrete-Fourier transform (DFT) at sampling frequency 16KHz. Initially, noise spectrum is estimated from the valleys of the amplitude spectrum [1]. The algorithm for noise estimation is as follows:
Compute the RMS value Y_{rms} of the amplitude spectrum $Y(k)$. We detect the minima of $Y(k)$ by obtaining the vector k_{min} such that $Y(k_{min})$ are the minima in $Y(k)$. Then the interpolation is performed between adjoining minima positions to obtain $Y_{min}(k)$ representing the minimum value sequences (MVS). We smooth the sequences by taking partial average called smoothed minimum value sequences (SMVS). An estimation of noise from the SMVS is survived by an overestimation and underestimation of the SNR which is controlled by proposed EDON. The block diagram of the noise estimation process is shown in Figure 1.

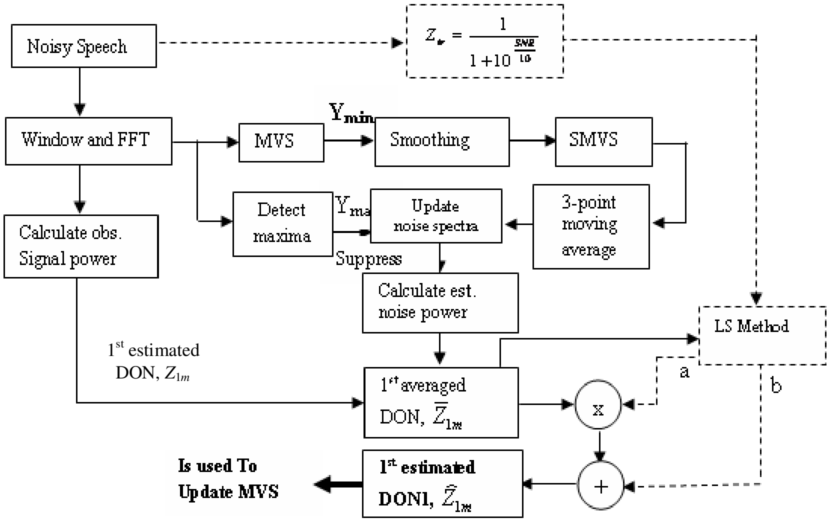
Noisy Speech
$Z_{tr} = \frac{1}{1+10^{\frac{SNR}{10}}}$
Window and FFT
MVS
Y_{min}
Smoothing
SMVS
Detect maxima
Y_{ma}
Update noise spectra
3-point moving average
Suppress
Calculate obs. Signal power
Calculate est. noise power
LS Method
1st estimated DON, Z_{1m}
1st averaged DON, $\bar{Z}_{1m}$
a
b
x
+
Is used To Update MVS
1st estimated DON1, $\hat{Z}_{1m}$

FIGURE 1: Block diagram of the 1st estimated DON, Z_{1m}.

2.2 Estimation of the Degree Of Noise (EDON)

In a single-channel method, we only know the power of the observed signal. To obtain EDON, we estimate noise of the observed signal in every analysis frame *m*. First white noise of various SNR is added to voiced vowel sounds. Now for each SNR, DON of each phoneme is estimated and averaged which corresponds the input SNR. Then each of these estimated 1st averaged DONs of each frame *m* for corresponding input SNR expressed as $\bar{Z}_{1m}$. The estimated $\bar{Z}_{1m}$ is aligned with the true DON (Z_{tr}) using the least-square (LS) method results the 1st estimated DON $Z_{1m}$of that frame. The true DON (Z_{tr}) is given by

$$Z_{tr} = \frac{P_d}{P_s + P_d} = \frac{1}{1+10^{\frac{dB}{10}}} \tag{1}$$

where *dB* is input SNR. The 1st averaged DON is

$$\bar{Z}_{1m} = \frac{1}{M}\sum_{m=1}^{M}\frac{P_{\eta}(m)}{P_{obs}(m)} \tag{2}$$

where, *M* are the noise added frames; $P_{\eta}(m)$and $P_{obs}(m)$ are the powers of noise and observed signals, respectively. Here it obvious that we consider only the voiced phonemes in our experiment. So the value of $\bar{Z}_{1m}$should be limited to voiced portion of a speech sentence. We used the same experiment with unvoiced speech. Practically the unvoiced portion contaminated with higher degree of noise. Hence the estimated noise is higher for unvoiced frame than from voiced frame. Consequently higher DON value is obtained from unvoiced frame than from voiced frame that is logically resemblance. The degree of noise estimated from a function using least square method is given as

$$Z_{tr} = a \times \bar{Z}_{1m} + b$$

here *a* and *b* are unknown. We estimate *a* and *b* via LS method, yielding $\bar{a}$ and $\bar{b}$and the estimated degree of noise is given by

$$Z_{1m} = \bar{a} \times \bar{Z}_{1m} + \bar{b} \tag{3}$$

where Z_{1m} is the 1st estimated DON of frame *m*. The value os Z_{1m} is applied to update the MVS. Next, the noise level is re-estimated and updated with the help of Z_{1m}. Finally, from the estimated noise, we again estimate 2nd averaged DON ($\bar{Z}_{2m}$) and similarly the 2nd estimated DON (Z_{2m}) which is used to estimate the noise weight for non linear weighted noise subtraction.

2.3 Noise Spectrum Estimation

We detect the minima $Y_{\min}(k_{\min}) \leftarrow \min(Y(k))$ values of amplitude spectrum Y(*k*) when the following condition (Y(*k*)<Y(*k*-1) and Y(*k*)<Y(*k*+1) and Y(*k*)<Y_{rms}) is satisfied. The $k_{\min}$ expresses the positions of the frequency bin index of minima values. Then interpolate between adjoining minima positions $(k_{\min} \leftarrow k)$ to obtain the minima value sequence (MVS) $Y_{\min}(k)$. Now we smooth the sequences by taking partial average called smoothed minima value sequence (SMVS). This process continuously updates the estimation of noise among every analysis frames. Now the noise spectrum is estimated from the SMVS and 1st estimated DON according to the condition

$$D_m(k) = Y_{\min}(k) + \left(\sqrt{Z_{1m}} \times Y_{rms}\right) \tag{4}$$

where Y_{rms} is the rms value of the amplitude spectrum. Then we made some updates of $D_m(k)$, the updated spectrum is again smoothed by three point moving average, and lastly the main maximum of the spectrum is identified and are suppressed [1]. Figure 2 shows the spectrums.

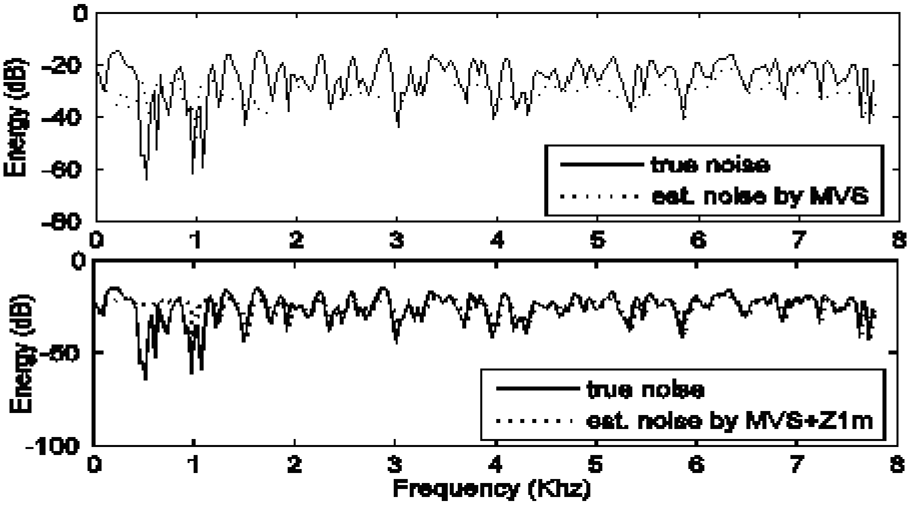

FIGURE 2: Noise spectrums (true and estimated).

2.4 Non-linear Weighted Noise Subtraction (NWNS)

Noise reduction in the front-end is based on implementation of the traditional spectral subtraction (SS) require an available estimation of the embedded noise, here, in time domain we named noise subtraction (NS). The goal of this section is to modify the noise subtraction process by adopting a non linear weight for minimizing the effect of residual noise in the processed speech and then to improve the performance by using EMD.

For subtraction in time domain, the estimated noise in the previous section is recombined with the phase of the noisy speech and inverse transformed one. Then we obtain $\hat{d}_{ss}(n)$ by withdrawing the effect of the window. The NWNS is given by:

$$s_1(n) = y(n) - \sqrt{\alpha \times Z_{tr}} \times \hat{d}_{ss}(n) \tag{5}$$

where $\alpha = 0.3019 + 6.4021 \times Z_{2m} - 14.109 \times Z_{2m}^2 + 9.8273 \times Z_{2m}^3$ is nonlinear weighting factor. We use least-square method for the estimation process. We find that for each input SNR, certain weight is required for best noise reduction results over wide ranges of SNR. In this experiment, we used 7 male and 7 female speakers of 10 different sentences at different SNR levels, randomly selected from the TIMIT database. We use 3rd degree polynomials to derive the above formulation. It is observed from Eq. (1) that it needs the input SNR. The input SNR can be estimated using variance is given by

$$SNR_{input} = 10 \log_{10} \left(\frac{\sigma_s^2}{\sigma_\eta^2} \right) \tag{6}$$

where, σ_s^2 and σ_η^2 are the variances of speech and noise respectively. We assume that due to the independency of noise and speech, the variance of the noisy speech is equal to the sum of the speech variance and noise variance. It is found that by adopting nonlinear weighted in NS, a good noise reduction is obtained. Although with the NWNS, we find the good performance with less musical noise by informal listening test but for further enhancement we cascade another method EMD and get better results.

3. CASCADE OF NWNS AND EMD

The general block diagram of the proposed system is shown in Figure 3. In the block diagram, first stage is incorporated a Noise Subtraction (NS) method with weight and second stage a Empirical Mode Decomposition (EMD) based adaptive thresholding method.

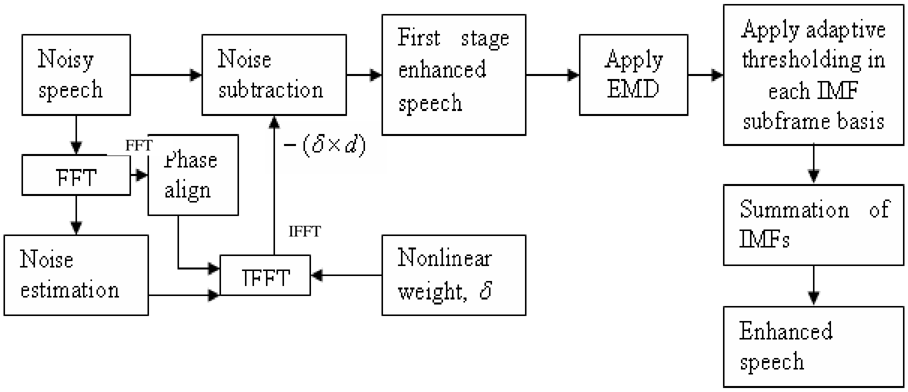

FIGURE 3: The block diagram of the two-stage NWNS+EMD method.

3.1 Empirical Mode Decomposition (EMD)

The principle of EMD technique is to decompose any signal $y(n)$ into a set of band-limited functions, which are the zero mean oscillating components, called simply the intrinsic mode functions (IMFs) [9]. Although a mathematical model has not been developed yet, different methods for computing EMD have been proposed after its introduction [15]. The very first algorithm, called as the sifting process, is adopted here to find the IMF's include the following steps;

1. Identify the extrema of $y(n)$
2. Generate the upper and lower envelopes ($u(n)$ and $l(n)$) by connecting the maxima and minima points by interpolation
3. Calculate the local mean $\mu_1(n)=[u(n)+l(n)]/2$
4. Since IMF should have zero local mean, subtract out $\mu_1(n)$ from $y(t)$ to obtain $h_1(t)$
5. Check whether $h_1(t)$ is an IMF or not
6. If not, use $h_1(t)$ as the new data and repeat steps 1 to 6 until ending up with an IMF.

Once the first IMF is derived, we should continue with finding the remaining IMFs. For this purpose, we should subtract the first IMF $c_1(n)$ from the original data to get the residue signal $r_1(t)$. The residue now contains the information about the components of longer periods. We should treat this as the new data and repeat the steps 1 to 6 until we find the second IMF.

3.2 Soft-thresholding

The soft thresholding strategy proposed in [16] for a frame, *m* of length *L* in transform-domain as

$$\widehat{Y}_q = \begin{cases} Y_q, & if \quad \phi \geq \sigma_n^2 \\ sign(Y_q)\left[\max\{0,(|Y_q| - j\gamma)\}\right], & otherwwise \end{cases} \tag{7}$$

where $\phi = \frac{1}{L}\sum_{q=1}^{L}|Y_q|^2$ denotes the average power of the frame, and σ_n^2 is the global noise variance of the speech, Y_q is qth coefficient of the frame obtained by the required transformation and $\widehat{Y}_q$ denotes to the thresholded samples of the frame. The multiplication factor $j\gamma$ is the linear threshold function while j being the sorted index-number of $|Y_q|$. An estimated value of γ can be obtained as:

$$\gamma = \frac{\lambda\sigma_n}{\sqrt{\frac{1}{Q}\sum_{q=1}^{Q} q^2}} \tag{8}$$

where λ is an adaptation factor and its value is determined experimentally such that $0<\lambda<1$. It is observed that the first part of Eq. (7) is for signal dominant frame when the condition satisfies, and second part is for noise dominant frame where soft thresholding will have to apply. So the classification of frames either to be signal dominant or noise dominant depends on average power of a frame and global noise variance of the given noisy speech. In this paper, we apply this soft thresholding strategy adaptively in each IMF, as discuss in the next section.

3.3 Adaptive thresholding

Soft thresholding strategy performs better on wide range of input SNR due to thresholded noise dominant frames only and kept remain the same in case of signal dominant frames but the misclassification of frames is a major drawback that causes musical noise [9]. Therefore this method is mainly appropriate for white noise. All the drawbacks can be significantly reduced with the proposed EMD based adaptive thresholding strategy with some modification of frame classification criteria. Since the IMFs will have different noise and speech energy distribution, so it suggests that each IMF will have a different noise and speech variance. After applying EMD, the soft thresholding technique is applied on each sub-frame of each IMF based on the computed variances. It is obvious that the variances will be changed for different sub-frames as well as with the individual IMF. The threshold will also be changed with newly computed variances and hence this technique is termed as adaptive thresholding. The proposed EMD based adaptive thresholding strategy for r^{th} subframe of $(i')^{th}$ IMF as:

$$\hat{Y}_{q,i'}^{(r)} = \begin{cases} Y_{q,i'}^{(r)}, & if\, \varphi_{i'}^{(r)} \geq 2\sigma_{n,i'}^2 \\ sign(Y_{q,i'}^{(r)})\left[\max\{0,(|Y_{q,i'}^{(r)}| - j'\hat{\gamma})\}\right], & otherwise \end{cases} \tag{9}$$

Here, $\hat{Y}_{q,i'}^{(r)}$ denotes to the thresholded samples of r^{th} subframe of the $(i')^{th}$ IMF, $Y_{q,i'}^{(r)}$ is q^{th} coefficient of r^{th} subframe of $(i')^{th}$ IMF and the multiplication $j'\hat{\gamma}$ is the adaptive threshold function while j' being the sorted index-number of $|Y_{q,i'}^{(r)}|$. The threshold factor $\hat{\gamma}$ is varied adaptively for individual IMF according to its variance. An estimated value of $\hat{\gamma}$ can be obtained as:

$$\hat{\gamma} = \frac{\sigma_{-n,i'}}{\sqrt{\frac{1}{Q}\sum_{q=1}^{Q} q^2}} \quad \text{or,} \quad \hat{\gamma} = \frac{\lambda\sigma_{n,i'}}{\sqrt{\frac{1}{Q}\sum_{q=1}^{Q} q^2}}$$

where, $Q = 64$, $\sigma_{-n,i'} = \lambda\sigma_{n,i'}$, $\lambda =$ adaptation factor and $\sigma^2_{n,i'} =$ noise variance of the $(i')^{th}$ IMF. Since global noise variance is estimated from silent frames, therefore, it assumes each frame as well as subframe belong that variance. That is why; the boundary for the classification of subframes should be set to two times of the globally estimated noise variance when noise variance and speech variance of that subframe are same. The enhanced speech signal of the EMD based adaptive thresholding is given by

$$s_2(n) = \sum_{i'=1}^{I}\left[\sum_{r=1}^{R}\left(\sum_{q=1}^{Q}\hat{Y}^{(r)}_{q,i'}\right)\right] \tag{10}$$

where, I=total number of IMFs,
R=total number of subframe and
Q=length of a subframe.

4. EXPERIMENTAL RESULTS AND DISCUSSION

We study the effectiveness of the proposed NWNS+EMD based adaptive thresholding algorithm are tested on the speech data corrupted by three different types of additive noise like white, pink and HF channel noise are taken from NOISEX database. N=56320 samples of the clean speech /she had your dark suit in greasy wash water all year/ from TIMIT database were used for all simulations. The noises are added to the clean speeches at different SNRs from –10dB to 30dB of step 5 to obtain noisy speech signals.

For evaluating the performance of the method, we are used the overall output and average segmental SNRs that are graphically represented as for measuring objective speech quality. The results of the average output SNR obtained from for white noise, pink noise and HF channel noise at various SNR levels are given in Table 1 for pre-processed speech in the first stage and final enhanced speech in the second stage respectively. Since in the real world environments, the noise power is sometimes equal to or greater than the signal power or the noise spectral characteristics sometimes change rapidly with time, NS or NWNS is not so effective in such situations. Because, there have to introduced large errors in the noise estimation process. EMD based adaptive thresholding method plays a vital role for the above case as found in Table 1. Table 2 presents a comparison the overall average output SNR among our previous method WNS and WNS+BSS with proposed method NWNS+EMD.

Input SNR	White noise		HF channel noise		Pink noise	
	NWNS	EMD	NWNS	EMD	NWNS	EMD
-10dB	-1.57	2.06	-7.47	-0.58	-7.06	-6.69
-5dB	2.39	5.69	-2.66	3.03	-2.32	-1.92
0dB	5.26	8.85	1.91	6.29	2.14	2.82
5dB	8.66	11.94	6.42	9.74	6.33	7.22
10dB	11.64	15.15	10.77	13.46	10.73	11.71
15dB	15.77	18.72	15.42	17.42	15.40	16.26
20dB	20.37	22.62	20.22	21.64	20.22	20.91
25dB	25.17	26.85	25.11	26.12	25.11	25.64
30dB	30.05	31.27	30.02	30.77	30.02	30.44

TABLE 1: The average output SNR for various types of noises at different input SNR by NWNS and NWNS+EMD (indicated as EMD).

Input SNR	White noise			HF channel noise			Pink noise		
	WNS	WNS+BSS	EMD	WNS	WNS+BSS	EMD	WNS	WNS+BSS	EMD
0dB	0.66	8.1	8.9	0.4	4.3	6.3	0.4	2.1	2.8
5dB	6.0	10.2	11.9	5.5	7.8	9.7	5.5	6.8	7.2
10dB	11.1	11.2	15.2	10.5	10.9	13.5	10.4	10.2	11.7
15dB	15.7	13.8	18.7	15.1	13.1	17.4	15.0	13.2	16.3
20dB	19.2	15.2	22.6	18.6	14.9	21.6	18.8	15.1	10.1
25dB	21.3	15.7	26.9	20.8	15.7	26.1	21.4	15.8	25.6
30dB	22.3	16.0	31.3	21.8	15.8	30.8	22.7	16.1	30.5

TABLE 2: The average output SNR for various types of noises at different input SNR by WNS, WNS+BSS (previous methods) and NWNS+EMD (indicated as EMD).

In terms of speech quality and intelligibility, the proposed two-stage (NWNS+EMD based adaptive thresholding method has to given a better tradeoff between noise reduction and speech distortion. We investigate this effect from the enhanced speech waveforms obtained from various methods as shown in Figure 4. It is observed from the waveforms that the enhanced speech is distorted in low voiced parts due to remove the noise in NS method whereas NWNS does not. A little amount of noise is removed from the corrupted speech by NWNS method. So in NS method there is a loss of speech intelligibility while NWNS maintains it. Although the EMD based adaptive thresholding can be able to successfully remove the noise from voiced parts but there is some noise remaining in the silent parts because of misclassification of subframes as signal-dominant. This remedy can be avoided using the proposed method. We also observed that by NS+EMD based adaptive thresholding method, there is loss of information in lower voiced parts and as a result speech intelligibility reduced. Moreover, the wavefrom obtained by NWNS+EMD based adaptive thresholding, it can be seen that there is no loss of information in lower voiced parts and maintains the speech intelligibility. We use two perceptually motivated objective speech quality assessments, namely the average segmental SNR (ASEGSNR) and the Perceptual Evaluation of Speech Quality (PESQ) to study the effectiveness of the proposed method. In Figures 5 and 6, it is observed that our proposed NWNS+EMD based adaptive thresholding approach achieve comparable improvements of speech quality. The PESQ scores of the speech at –10dB and –5dB (pink and HF channel noise) are almost equal to input PESQ scores. This is due to the presence of musical noise in first stage

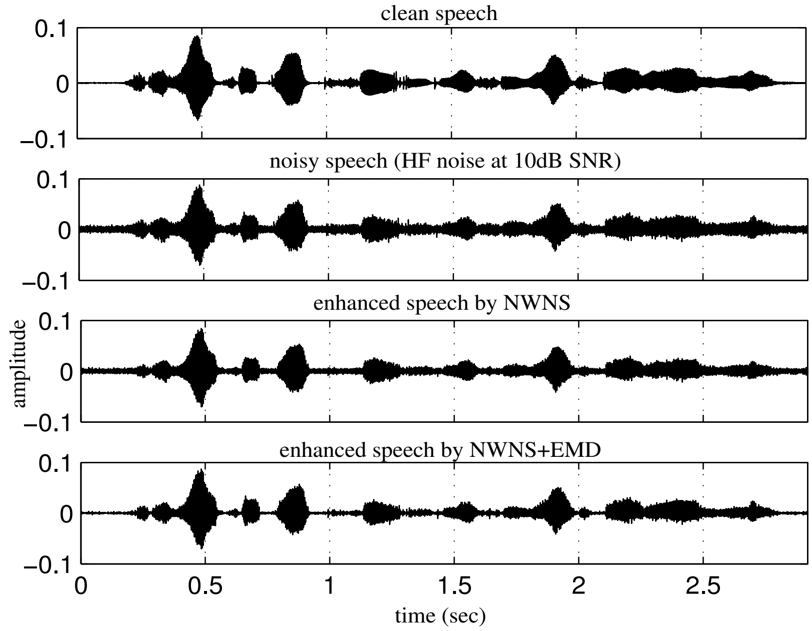

FIGURE 4: Speech waveforms of (from top) clean, noisy (HF noise at 10dB), enhanced by NWNS and NWNS+EMD.

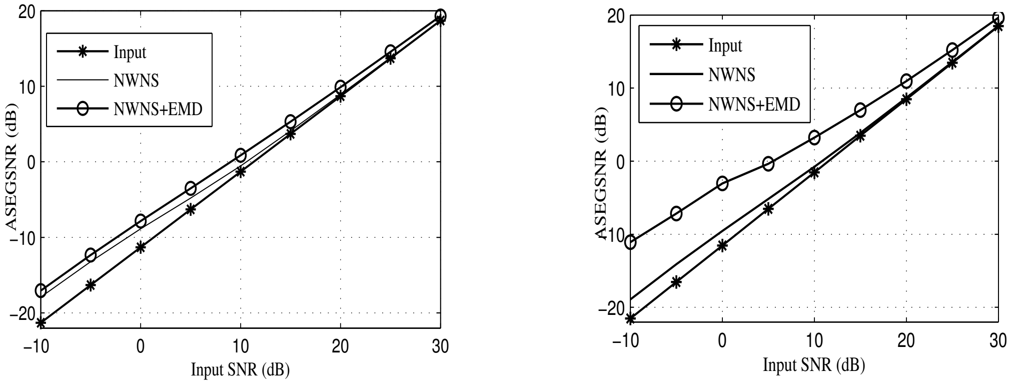

FIGURE 5: Comparisons of the average output segmental SNR (ASEGSNR) by NWNS and NWNS+EMD methods for pink noise (left) and HF channel noise (right).

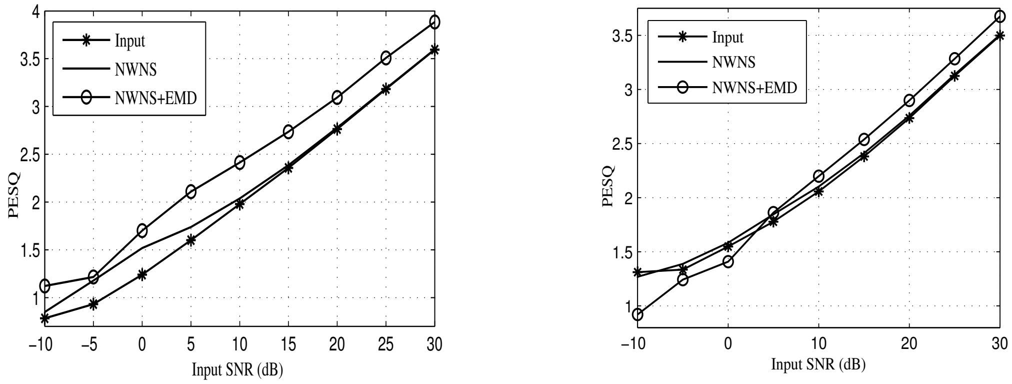

FIGURE 6: Comparison of PESQ scores by NWNS and NWNS+EMD methods for pink noise (left) and HF channel noise (right).

5. CONCLUSION & FUTURE WORK

In this paper, we presented a new algorithm to effectively remove the noise components in all frequency levels of a noisy speech signal. Our aimed to improve SNR of noise contaminated speech by removing and/or reducing noise using a two-stage processing technique; namely, a time domain nonlinear weighted noise subtraction (NWNS) followed by an Empirical Mode Decomposition (EMD) based adaptive thresholding. The first enhanced speech became as input of the second stage for further enhancement and obtained final enhanced speech after second stage processing. We introduced the degree of noise (DON1 and DON2) estimation process. DON1 was used to improve noise estimation accuracy and DON2 to calculate nonlinear weighting factor for NWNS in order to reduce musical noise. The parameters of DON1 and DON2 were estimated for white noise and we used the same parameters for all color/real world noises. Since the empirical mode decomposition (EMD) was fully data adaptive and highly effective for nonlinear and nonstationary data, it overcame inadequacy effect of the first stage for assumption as stationary of nonstationary speech segment. We combined NWNS+EMD based adaptive thresholding enhancement algorithm which worked most efficiently for wide range of input SNR. It was found that the amount of this improvement decreased when the interfering source power was minimal. This was because the algorithm was dependent upon the interfering noise signal estimation in the first stage and also dependent upon the adaptation factor and adaptive threshold factor in the second stage. When the interfering noise power was increased (up to 0dB), the proposed methods were able to perform better noise estimation. However, as the interfering noise power became much larger, as was true for extremely small SNR's (<0dB), the algorithm did not perform well in the case of color noises due to the inability of the method to obtain an adequate estimate of the original signal. The performance of the proposed method over speech contaminating with white noise or color noise was good based on objective measures and spectrograms and waveforms analysis.

Since in single channel speech enhancement method, there was difficulty removing all the noise components from speech without introducing musical noises or distortions, hence in this regard further research can be conducted to increase the accuracy of noise estimation (DON1) and also the more adjustment needed of the nonlinear weight (DON2) for voiced/unvoiced sections for underlying noisy speech to reduce musical noise and to improve speech quality. All EMD based algorithm suffers from computational complexity and the empirical process takes long time and is not applicable for real time processing. Therefore, it is suggested that more research can be conducted on insight the EMD making it less empirical and more mathematical.

6. REFERENCES

1. M. E. Hamid, K. Ogawa, and T. Fukabayashi, "*Improved Single-channel Noise Reduction Method of Speech by Blind Source Separation*", Acoust. Sci. & Tech., Japan, 28(3):153-164, 2007

2. J. Benesty, S. Makino, and J. Chen, "*Speech Enhancement*", Springer-Verlag Berlin Heidelberg, 2005

3. M. M. Sondhi, C. E. Schmidt and L. R. Rabiner, "*Improving the Quality of a Noisy Speech Signal*", Bell Syst. Techn. J., vol. 60, October 1981

4. 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, April 1979

5. R. Martin, "*Spectral Subtraction Based on Minimum Statistics*", Proc. EUSIPCO, pp. 1182-1185, 1994
6. R. Martin, "*Speech Enhancement based on Minimum Mean-Square Error Estimation and Supergaussian Priors*", IEEE Trans. Speech and Audio Process., vol. 13, no. 5, pp. 845-858, Sept. 2005

7. C. He, and G. Zweig, "*Adaptive two-band spectral subtraction with multi-window spectral estimation*", ICASSP, vol. 2, pp. 793-796, 1999

8. S. C. Liu, "*An approach to time-varying spectral analysis*", J. EM. Div. ASCE 98, 245-253, 1973

9. N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shin, Q. Zheng, N. C. Yen, C. C. Tung, and H. H. Liu, "*The Empirical Mode Decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis*", Proceeding Royal Society London A, vol. 454, pp. 903-995, 1998

10. S. F. Boll, and D. C. Pulsipher, "*Suppression of Acoustic Noise in Speech using Two-Microphone Adaptive Noise Cancellation*", Correspondence, IEEE Transactions on Acoustics, Speech and Signal Processing, vol. ASSP-28, no. 6, pp. 752-753, Dec 1980

11. Z. Xiaojie, L. Xueyao, Z. Rabu, "*Speech Enhancement Based on Hilbert-Huang Transform Theory*", in First International Multi-Symposiums on Computer and Computational Sciences, pp. 208-213, 2006

12. P. Flandrin, P. Goncalves and G. Rilling, "*Detrending and Denoising with Empirical Mode Decompositions*", In Proc., EUSIPCO, pp.1581-1584, 2004

13. K. Khaldi, A. O. Boudraa, A. Bouchikhi, and M. T. H. Alouane, "*Speech Enhancement via EMD*", in EURASIP Journal on Advances in Signal Processing, vol. 2008, Article ID 873204, 8 pages, 2008

14. T. Hasan, and M. K. Hasan, "*Suppression of Residual Noise from Speech Signals using Empirical Mode Decomposition*", Signal Processing Letters, IEEE, vol. 16, no. 1, pp. 2- 5, Jan 2009

15. X. Zou, X. Li, and R. Zhang, "*Speech Enhancement Based on Hilbert-Huang Transform Theory",* First International Multi-Symposiums on Computer and Computational Sciences, 1: 208–213, 2006

16. Flandrin, P., Rilling, G. and Goncalves, P., "Empirical mode decomposition as a filter bank," IEEE Signal Processing Letters, 11(2), pp. 112-114, 2004