SindhUniv. Res. Jour. (Sci. Ser.) Vol. 47 (3) 519-522 (2015)

SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

Enhancement of Speech Signals Using Multiple Statistical Models

R. A. KHALIL, S. M. ASHRAF, T. JAN[++], A. JEHANGIR, J. B. KHAN

Department of Electrical Engineering, University of Engineering & Technology Peshawar, Pakistan

Received 11th November 2014 and Revised 6th May 2015

Abstract: This paper proposed a correlative analysis of two recent approaches for noise reduction based on minimum mean-square error (MMSE) and hard thresholding or shrinkage for a provided speech signal. Based on designed algorithm, both noise reduction techniques are real-time data driven schemes. The corrupted speech signal is disintegrate into vacillate frames by a well know sifting process. The fundamental principle of these two techniques is the reconstruction of signal after removal of the noisy signal frames that are already filtered, applying the Minimum Mean-Square Error (MMSE) filtering technique, and hard threshold utilizing shrinkage. These techniques are investigated and correlated with standard denoising methods using different evaluation methods such as, signal-to-noise ratio (SNR) and signal-to-distortion ratio (SDR). The literature of these schemes is retrained to noisy signals that are corrupted by white noise. Experimental results attained demonstrate that proposed MMSE filtering scheme for denoising perform better than hard thresholding or shrinkage in both the segmental and overall consideration of noisy corrupted speech signals.

Keywords: Minimum Mean-Square Error (MMSE), Hard Thresholding, Additive White Gaussian Noise (AWGN), Automatic Speech Recognition (ASR), Signal-to-Noise Ratio (SNR)

1. INTRODUCTION

In automatic speech recognition systems (ASR), background noise causes distortion and disruption. Noise such as additive white Gaussian noise (AWGN) is one of the major elements that cause performance degradation in ASR systems due to its imitative effect of randomness. Enhancement of speech signals is a restrained complication in signal processing, specifically with the understanding of AWGN and various noise reduction techniques have been applied to tackle noisy signals (Khaldi *et al.*, 2008). Linear filtering provides accurate results when noise can be estimated in a given distorted signal. However, in case of signal consisting tipped edges or signals with short period these methods are not so adequate. To overcome these demerits and imperfections, techniques with non linearity have been presented based on thresholding of wavelets. It states that magnitudes of signal are greater than magnitudes of noise in a selected wavelet. If the magnitudes of selected wavelet are less than a predefined threshold then its coefficients are reduced to zero. The main limitation of wavelet approach is its base functions. These base functions are usually steady in nature, and are not applicable to all real-time signals. In order to avoid this limitation a scheme known as time-frequency decomposition is used. Mostly the wavelet packets represent a larger number of real-time signals (Khaldi *et al.*, 2008). Despite this, for each analysis the base function such as Gabor functions, undamped or damped sinusoidal functions must be defined.

In (Fenghua *et al.*, 2010) the method introduced depends upon wavelet packets MMSE filtering techniques. Voice activity detection (VAD) is used to remove noise components from corrupted noisy speech. The experimental results obtained provide sufficient noise reduction without considerable distortion in speech signals.

In (Dang and Nakai, 2011; Rao *et al.*, 2011), an adaptive filtering has been applied to restrain the interfering signals and reduce noise effect in the given speech signal. Wiener filtering with some modification has been used to denoise the speech signal distorted by fluctuating noise. The conventional methods used for speech enhancement takes average of the noise power spectrum as the signal-to-noise ratio (SNR) of signal ranges from 5dB to 10dB. It then tracks the spectrum of noise directly from speech signal and revised it at each frame. At low fluctuating signal-to-noise ratio, this method results in a better performance over classical MMSE algorithms**.**

In (Ephraim and Malah, 1984; Rao *et al.* 2011; Ykhlef and Ykhlef, 2014), methods for speech enhancement have been discussed based on Minimum Mean Square Error short-time Log spectral amplitude (MMSE-LSA) and short-time spectral amplitude (STSA) estimator. Single channel speech enhancement algorithm is used with a multiplicative gain filter in frequency domain. The spectral gain with short-term spectrum of corrupted speech signal results in an

[++]Corresponding Author: Tariqullah Jan, tariqullahjan@uetpeshawar.edu.pk, +92-91-9216498,
Department of Electrical Engineering, University of Engineering & Technology Peshawar, Pakistan.

estimated clean speech spectrum. In order to reconstruct the enhanced signal, the STSA of the speech signal is approximated, and combined with the short-time phase of the corrupted speech signal (Rao *et al.* 2011). The spectral subtraction algorithm and Weiner filtering are well recognized models of linear filtering used to de-noised corrupted speech signals. The spectral subtraction STSA estimator is derived from variance estimator while the Weiner STSA estimator is derived from MMSE spectral signal estimator. These estimators are amplitude based and not optimal under the assumed criterion and statistical model.

In this paper, two noise reduction techniques have been evaluated for the enhancement of speech signal. The first technique uses the minimum mean-squared error (MMSE) filtering scheme (Khaldi *et al.*, 2008), while the second one uses hard thresholding or shrinkage (Khaldi *et al.*, 2008). These techniques are applied and tested on given corrupted speech signals with various noise levels, followed by comparative analysis using different evaluation schemes such as, signal-to-noise-ratio (SNR) and signal-to-distortion-ratio (SDR). Experimental results clearly elaborate the performance of these two methods for the noisy speech signals.

2. MATERIAL AND METHODS

Proposed Method

The proposed speech enhancement system is depicted in Figure-1: In this figure, $s(t)$ is a clean speech signal, followed by a distorted and noisy version $z(t)$. Then windowing technique $w(t)$ has been applied to $z(t)$ to obtain the frames $\mathrm{W}_j(t)$. Now the denoising principle will be discussed in the forthcoming section to elaborate the basic concept of the current work.

2.1 Principle of Denoising

Initially a clean speech signal $s(t)$ is distorted by noise such as additive white Gaussian (AWGN) noise $b(t)$ as follows (Khaldi *et al.*, 2008):

$$z(t) = s(t) + b(t) \qquad \text{(i)}$$

The noise corrupted speech signal is disintegrate into vacillate windows $\mathrm{W}_j(t)$ using sifting process, such that

$$z(t) = \sum_{j=1}^{c} \mathrm{W}_j(t) + r_c(t) \qquad \text{(ii)}$$

The estimation of $\widetilde{f}_j(t)$ is given by $f_j(t)$ and it is based on consideration of noise level in the given signal windows $\mathrm{W}_j(t)$ is depicted by (Khaldi *et al.*, 2008):

$$\widetilde{f}_j(t) = \Gamma\left[\mathrm{W}_j(t);\ \tau_j\right] \qquad \text{(iii)}$$

where $\Gamma[\mathrm{W}_j(t);\ \tau_j]$ is a pre-processing base function, elaborated by a set of parameters τ_j also known as threshold parameter. These parameters are tested on the temporal frames $\mathrm{W}_j(t)$ of the signal. The base function Γ depends upon the estimation of level of noise in the given speech signal. When this noise is estimated correctly, the base function Γ is scaled down to MMSE filtering. However, if the noise is hard to estimate, the pre-processing can be hard thresholding or shrinkage. Finally, the de-noised signal, $\tilde{z}(t)$, is given by (Khaldi *et al.*, 2008):

$$\tilde{z}(t) = \sum_{j=1}^{c} \widetilde{f}_j(t) + r_c(t) \qquad \text{(iv)}$$

2.2 MMSE Filtering Technique

Estimation of speech noise will be performed by using Boll's method (Boll, 1979). According to, the silence intervals of the given speech signal are detected and examined. After detection of these intervals power spectral analysis is performed to estimate the noise. The silence moments of the noisy speech signal are expressed by taking the average of the power spectra. It is then applied on the first *M* number of temporal frames; the relative expression is given by (Khaldi *et al.*, 2008):

$$\left|\widetilde{\mathbf{V}}(\mathrm{fs}, m)\right|^2 = \frac{1}{M}\sum_{i=0}^{M-1} |\mathbf{V}(\mathrm{fs}, i)|^2 \qquad \text{(v)}$$

where $\mathbf{V}$ is the initial silence interval of the given speech signal and $|\mathbf{V}(\mathrm{fs}, i)|$ is the value of power spectra of frame i at certain discrete frequency fs. This technique results in accurate noise estimation in the provided signal. Experimental results obtained after decomposition of the provided speech signal, the first frame corresponds to the respective silence sequence. Hence, the first frame can be used for correct estimation of level of noise. According to, the noise level follows the very first frame (k = 1) which can be expressed by (Khaldi *et al.*, 2008) **(Fig-1).**

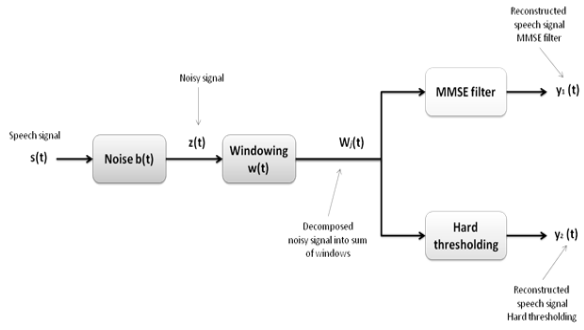


Fig-1: Block Diagram of Proposed Method

$$\tilde{\sigma}_k = \frac{\tilde{\sigma}_1}{\sqrt{2}^{k-1}} \quad \text{with } k > 1, \qquad \text{(vi)}$$

where $\tilde{\sigma}_1$ is the level of noise of first frame.

The MMSE strategy will be used to de-noise the signal. Thus, filtering of each temporal frame is done by MMSE filter and expressed as follows (Khaldi *et al.*, 2008):

$$\tilde{\mathrm{F}}_j(\mathrm{fs}, m) = \mathrm{H}(\mathrm{fs}, m)\mathrm{W}_j(\mathrm{fs}, m) \qquad \text{(vii)}$$

where $\mathrm{W}_j(\mathrm{fs}, m)$ are the noisy spectral frames and $\tilde{\mathrm{F}}_j(\mathrm{fs}, m)$ are the estimated noisy spectral frames. These m frames are examined at the given discrete frequency fs. $\mathrm{H}(\mathrm{fs}, m)$ is elaborated as follows (Khaldi *et al.*, 2008):

$$\mathrm{H}(\mathrm{fs}, m) = \frac{\mathrm{SNR}_{\mathrm{prio}}(\mathrm{fs}, m)}{1 + \mathrm{SNR}_{\mathrm{prio}}(\mathrm{fs}, m)} \qquad \text{(viii)}$$

The prior signal-to-noise $\mathrm{SNR}_{\mathrm{prio}}$ is based on estimation from previous frame. This estimation is based on $\tilde{\mathrm{F}}_j(\mathrm{fs}, m-1)$ from proceeding frames and estimation of instantaneous signal-to-noise ratio $\mathrm{SNR}_{\mathrm{inst}}$ (Khaldi *et al.*, 2008):

$$\mathrm{SNR}_{\mathrm{prio}}(\mathrm{fs}, m) = \alpha \frac{\mathrm{F}^2(\mathrm{fs}, m-1)}{\mathrm{S}^2(\mathrm{fs}, m-1)} + (1 - \alpha)\max(\mathrm{SNR}_{\mathrm{inst}}(\mathrm{fs}, m), 0) \qquad \text{(ix)}$$

where α is described as a weighting factor. The value of weighting factor is considered taken equal to 0.85. The $\mathrm{SNR}_{\mathrm{inst}}$ represent the instantaneous signal-to-noise ratio (SNR), which intern describes the prior estimation of signal-to-noise ratio $\mathrm{SNR}_{\mathrm{prio}}$ (Khaldi *et al.*, 2008):

$$\mathrm{SNR}_{\mathrm{inst}} = \frac{\mathrm{W}^2(\mathrm{fs}, m)}{\mathrm{V}^2(\mathrm{fs}, m)} \qquad \text{(x)}$$

2.3 Hard Thresholding Technique

Hard thresholding technique is applied to the input noisy signal frames $\mathrm{W}_j(t)$ before reconstructing the original smooth signal. In this scenario, the estimation of noise in the given signal can be calculated by threshold parameter τ expressed by following equation:

$$\tau = \sqrt{2\log(T)\,\sigma} \qquad \text{(xi)}$$

where T represents the length of the signal and σ is the scaled level of estimated noise. The $\tilde{\sigma}_1$ is given by (Khaldi *et al.*, 2008):

$$\tilde{\sigma}_1 = 1.4826 \times \mathrm{Median}\{|\mathrm{W}_1(t) - \mathrm{Median}\{\mathrm{W}_1(t)\}|\} \qquad \text{(xii)}$$

the level of noise $\tilde{\sigma}_k$ from (vi) of the signal will be estimated by applying $\tilde{\sigma}_1$.

There are various non-linear shrinkage functions. In the proposed work, hard shrinkage is used which will result in de-noising of the corrupted speech signal:

$$f_j = \begin{cases} \mathrm{W}_j(t) & \text{if } |\mathrm{W}_j(t)| > \tau_j, \\ 0 & \text{if } |\mathrm{W}_j(t)| \le \tau_j. \end{cases} \qquad \text{(xiii)}$$

The hard thresholding technique is also called as hard shrinkage.

3. RESULTS AND DISCUSSION

In this section, the two denoising techniques are applied on noisy speech signals corrupted by additive white Gaussian noise (AWGN) using different values of signal-to-noise ratios (SNRs). For experimental tests four clean and smooth speech signals were used, 2 male and 2 female. Each speech signal is sampled at a rate of 16 kHz. As indicated, the denoising techniques depend upon estimation of noise.

So, the MMSE filtering technique is used if the pre-speech region of the noisy speech signal is detected. Otherwise, hard thresholding or hard shrinkage is used for denoising.. More precisely, the signal-to-noise ratio (SNR) is employed to compare the MMSE filtering scheme and hard thresholding or shrinkage. The SNR is expressed as

$$\mathrm{SNR} = 10\log_{10}\frac{\sum_{i=1}^{T}(f(t_i))^2}{\sum_{i=1}^{T}(f(t_i) - \tilde{f}(t_i))^2} \qquad \text{(xvi)}$$

where $f(t_i)$ is the original speech signal and $\tilde{f}(t_i)$ is the denoised speech signal. As *T* is the length of speech signal. The output comparison has been carried out on the basis of ΔSNR. The ΔSNR is calculated as;

$$\Delta\mathrm{SNR} = m_{\mathrm{o}} - m_{\mathrm{i}} \qquad \text{(xvii)}$$

where m_o is the output SNR and m_i is the input SNR of the speech signal.

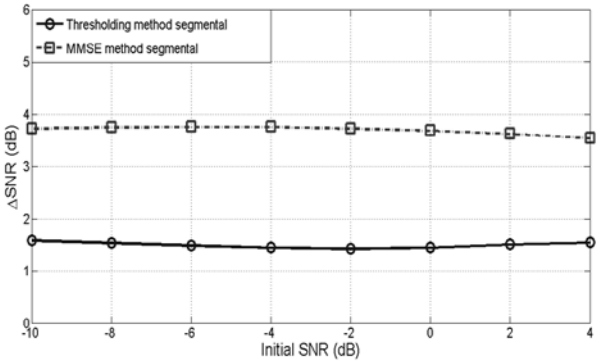


Fig-2: ΔSNR comparison on adaptive segment of signal for initial SNR ranges from -10 dB to 4 dB. Experimental results are average of 50 independent tests.

The experimental results obtained in the simulations for MMSE filtering and hard thresholding are compared on segment of a signal taken adaptively shown in **(Fig-2)** and overall signal shown in **(Fig-3).** The range of SNR is taken from -10dB to 4dB. For each initial SNR, total 50 independent random tests have been performed and average results were calculated as ΔSNR.

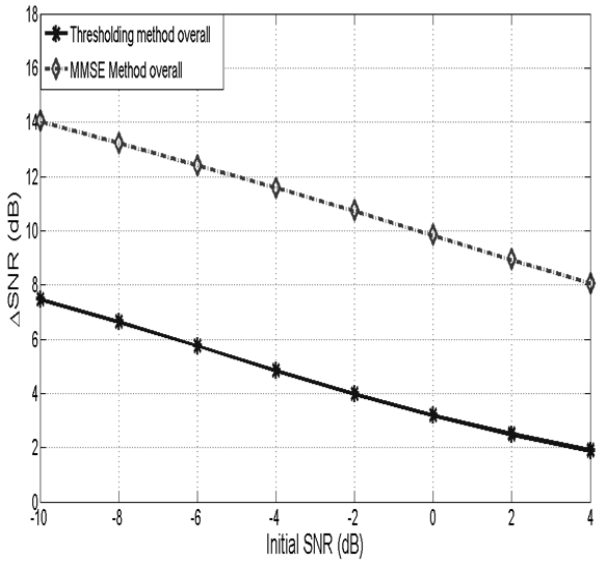


Fig-3: ΔSNR comparison on overall signal for initial SNR ranges from -10 dB to 4 dB. Experimental results are average of 50 independent tests.

From (**Fig-2)** and (**Fig-3**), we can observe that in the proposed scheme MMSE filtering technique offers improved denoising performance as compared to hard thresholding in term of SNR for segment of signal as well as overall signal.

4. CONCLUSION AND FUTURE WORK

This paper analyzed two new techniques for de-noising of the speech signals. Both techniques are based on the temporal frames and fully data-driven. The literature of these schemes is retrained to noisy signals that are corrupted by white noise. The performance evaluation of the two different techniques is investigated and evaluated using different evaluation methods such as, signal-to-noise-ratio (SNR). The experimental results attained demonstrate that MMSE filtering scheme for denoising perform better than the hard thresholding or shrinkage in both the segmental and overall consideration of noisy corrupted speech signals. Hence these techniques can be implemented in applications such as ASR systems for suppression of background noise, CDMA systems for cancellation of inter-channel interference, Cochlear implant (CI) to decrease the effect of noise and various applications in speech communications.

The reason for low performance of hard thresholding is that it is difficult to estimate the noise level for higher SNRs. Due to which it also degrade the speech signal after removal of noise. In order to minimize this degradation, a modified soft-thresholding technique can be introduced that works adaptively on a frame basis to denoised the corrupted speech signal in future.

REFERENCES:

Boll S. F. (1979) Suppression of acoustic noise in speech using spectral subtraction. IEEE Transactions on Acoustics, Speech and Signal Processing, vol.ASSP-27, 113-120.

Dang, X., and T. Nakai, (2011) Noise reduction using modified phase spectra and Wiener Filter. IEEE International Workshop on Machine Learning for Signal Processing (MLSP).

Donoho D. L. (1995) De-noising by soft-thresholding. IEEE Transactions on Information Theory, vol. 41, no. 3, 613–627.

Ephraim, Y. and D. Malah (1984) Speech enhancement using a minimum mean-square error short time spectral amplitude estimator. IEEE Transactions on Acoust., Speech, and Signal Process., vol. 32, no. 6, 1109–1121

Fenghua. Z., L. Yang, J. Wang and S. Qiang (2010) Speech signal enhancement through wavelet domain MMSE filtering. IEEE International Conference on Computer, Mechatronics, Control and Electronic Engineering (CMCE), Vol. 5, 24-26.

Khaldi. K., A. O. Boudraa, A. Bouchikhi and M. T. Alouane (2008) Speech enhancement via EMD. EURASIP J. Advaced Signal Processing, Article ID 873204.

Rao, C.V.R., M.B.R. Murthy, K. S. Rao (2011) Speech enhancement using perceptual Wiener filter combined with unvoiced speech — A new Scheme. IEEE international Conference on Recent Advances in Intelligent Computational Systems (RAICS) .

Rao, C.V.R., M.B.R. Murthy, K. S. Rao, K. A. Sheela (2013) A novel two stage single channel speech enhancement technique. IEEE Annual India Conference (INDICON) India.

Ykhlef, F. and H. Ykhlef (2014) A smoothed Minimum Mean-Square Error Log-Spectral Amplitude estimator for speech enhancement. IEEE International Conference on Multimedia Computing and Systems (ICMCS).