ELECTRONICS AND COMMUNICATION ENGINEERING DEPARTMENT NATIONAL INSTITUTE OF TECHNOLOGY SIKKIM-737139

A MINOR PROJECT
REPORT ON

Speech Enhancement using the knowledge of similar nature of speech segment

SUBMITTED ON-25/05/2019

SUBMITTED BY:

AKSHEETHA MUTHUNOORU
(B160032EC)

LOHITH SURISETTI
(B160014EC)

SHIVA KUMAR NAGULA
(B160090EC)

UNDER GUIDANCE OF:
Mr. AVINASH KUMAR
ASSISTANT PROFESSOR
ELECTRONICS AND COMMUNICATION ENGINEERING DEAPARTMENT
NATIONAL INSTITUTE OF TECHNOLOGY SIKKIM
RAVANGLA BARFUNG BLOCK, SOUTH SIKKIM-737139

CERTIFICATE

This is to certify that the project entitled **"Speech Enhancement using the knowledge of similar nature of speech segment"** is submitted by Aksheetha Muthunooru (B160032EC), Lohith Surisetti (B160014EC) and Shiva Kumar Nagula (B160090EC), in **"Electronics and Communication Engineering",** during session 2018-2019 at National Institute of Technology Sikkim. It is an authentic record of research work carried in ECED in NIT Sikkim.

The Report is based on the candidate's own work and has not submitted elsewhere.

Dr. Sanjay Kumar Jana
HOD (i/c), ECE Department

Mr. Avinash Kumar
Project Supervisor

ACKNOWLEDGEMENT

“Gratitude is not a thing of expression; it is more a matter of feeling.”

I would like to express my sincere gratitude to our guide **Mr. Avinash Kumar** for his vital support, guidance and encouragement, without which the completion of mini project would not have come forth. With his valuable suggestions and guidance it has been helpful in various phases of the mini project.

During the training period many people have best owed and supported me. Special thanks to **Dr. Sanjay Kumar Jana**, HOD(i/c), ECE Department, NIT Sikkim, faculty members & Ph.D. Scholars for their kind gesture and ever helping nature which made me in completion of mini project.

Finally, I would like to thank God Almighty for giving me the strength, knowledge, ability and opportunity to undertake this project study and to preserve and to complete it satisfactorily.

AKSHEETHA MUTHUNOORU **LOHITH SURISETTI** **SHIVA KUMAR**

INDEX

ABSTRACT

The aim of the speech enhancement is to improve the intelligibility and quality of the speech. By recording the speech signal in the noisy environment, clean speech signals are degraded. Speech enhancement [2] reduces the noise without distorting the original (clean) signal. In this work, a speech signal is enhanced based on non-local means (NLM)[1,2] estimation and variational mode decomposition (VMD)[1]. The NLM estimation is effective in removing noises whenever non-local similarities are present among the samples of the signal under consideration. However, it suffers from the issue of under-averaging in those regions where amplitude and frequency variations are abrupt. Since speech is a non-stationary signal, the magnitude and frequency vary over the time. Consequently, NLM is not that effective in removing the noise components from the speech signal as observed in the case of image enhancement. To address this issue, the noisy speech signal is first decomposed into variational mode functions (VMFs)[1] using VMD. Each of the VMFs represents a small portion of the overall frequency components of the signal. The VMFs are then combined into different groups depending on their similarities to reduce computational cost. Next, the non-local similarity present in each group of VMFs is exploited for an effective speech enhancement through NLM estimation. The enhancement performance of the proposed method is compared with the existing speech enhancement techniques. The experimental results presented in this study show that, the proposed method provides better performance.

1. INTRODUCTION

Speech enhancement aims to improve speech quality by using various algorithms. The objective of enhancement is improvement in intelligibility and/or overall perceptual quality degraded speech signal using audio signal processing techniques. Enhancing of speech degraded by noise, or noise reduction, is the most important field of speech enhancement, and used for many applications such as mobile phones, VoIP, teleconferencing systems, speech recognition, and hearing aids

The suppression of noise components from speech signal to improve the quality and intelligibility is not only essential but also extremely challenging. Over the years, several approaches for speech enhancement have been reported. Most of the classical speech enhancement approaches are subtractive in nature. Then, the estimate of the noise spectrum is subtracted from the noisy speech spectrum to enhance the signal quality. The performance of such approaches is highly dependent on the accuracy with which the non-speech region is detected and robust estimation of instantaneous noise spectrum. Several techniques have been proposed for estimating the noise spectrum from the noisy speech signal. However, such spectral enhancement methods introduce distortion in the enhanced speech signal due to deviations in estimated and actual instantaneous noise spectrum. In the enhancement approaches presented in, the high signal to noise ratio (SNR)[6] regions are identified and relatively more enhanced compared to the low SNR regions.

The non-local means estimation, a well explored method for denoising image and electrocardiography (ECG) signals, is effective in removing the noises whenever non-local similarities are present among the samples of the signal. Since speech is a non-stationary signal, the magnitude and frequency vary over the time. Consequently, NLM is not that effective in removing the noise components from the speech signal as observed in the case of image and

ECG enhancement. This issue can be addressed up to an extent by decomposing the signal into different narrow-band regions. The VMD algorithm decomposes a signal into a predefined number of narrow-band variational mode functions (VMFs). Each of the VMFs represents some smaller portion of the overall frequency band of the signal. Unlike the noisy speech signal, the VMFs do not have abrupt amplitude and frequency variations. Through this motivation, a speech enhancement approach is proposed in this report by utilizing the efficacy of VMD and NLM estimation.

2. LITRATURE SURVEY

In frequency-domain speech enhancement, the magnitude spectrum of the speech signal is estimated and combined with the short-time phase of the degraded speech to produce the enhanced signal. The spectral subtraction algorithms and Wiener filtering are well-known examples. In the spectral subtraction algorithms, the STSA of noisy and noise signals are estimated as the square root of maximum likelihood estimation of each signal spectral variance , and then subtracted from the magnitude spectrum of the noisy signal. In the Wiener filtering algorithm, the estimator is obtained by finding the optimal MMSE estimates of the complex Fourier transform coefficients. Both spectrum subtraction and Wiener filtering algorithms are derived under Gaussian assumption for each spectral component. Spectrum subtraction and Wiener filtering are not optimal spectral amplitude estimators under the assumed Gaussian model; the spectral amplitude estimators are more advantageous than spectral estimators from a perceptual point of view. Ephraimand Malah formulated an optimal spectral amplitude estimator, which, specifically, estimates the modulus (magnitude)of each complex Fourier coefficient of the speech signal in a given analysis frame from the noisy speech in that frame.

In order to derive the MMSE STSA estimator, the a priori probability distribution of the speech and noise Fourier expansion coefficients should be assumed since these are unknown in reality. Ephraim-Malah derived their spectral amplitude estimator based on a Gaussian model. They argued that this assumption utilizes asymptotic statistical properties of the Fourier coefficients. Specifically, according to , they assumed that the Fourier expansion coefficients of each process can be modeled as statistically independent Gaussian random variables, real- and imaginary parts of each component is independent to each other, and the mean of each coefficient is assumed to be zero and the variance time-varying. The assumption is motivated by the central limit theorem. Central limit theorem/asymptotic statistical properties hold more strongly when the analysis frame size is long, which somewhat conflicts with the "short-time" requirement. Porter and Boll , Brehm and Stammler , and Martin recognized that the DFT coefficients of clean speech derived from speech frames having a length of about the span of correlation within the signal are not normally distributed, and might be better modeled by a Gamma or a Laplacian distribution when

The DFT frame length is in the range of 10-40 ms, which is a typical frame size in speech application. The MMSE estimation of speech spectrum has received considerable research attention. Although significant progress has been made in the above mentioned methods, each of them has its own weaknesses and limitations either on the underlying assumptions or derivation of the estimators. The fact that Laplacian and Gamma are better than the Gaussian PDF for modeling the speech DFT coefficients has been recognized and empirically verified by many researchers. Therefore, many have focused on Laplacian/Gamma based estimators. None of them, however, has succeeded in finding the preferred optimal spectral amplitude estimator in the MMSE sense.

3. PROPOSED METHOD

Speech enhancement approach is proposed in this report by utilizing the efficacy of VMD and NLM estimation .The block diagram summarizing the proposed method for speech enhancement is shown in the following figure. In this approach, the speech enhancement is performed by processing the noisy speech signal through the following steps:

i) The noisy speech signal is decomposed into k number of VMFs using VMD. The VMFs having lower center frequency predominantly represents the high magnitude vowel-like regions where as the VMF having higher center frequency represent the unvoiced sound units.
ii) Then, the VMFs are divided into j groups depending on the similarity in their center frequencies and magnitude spectrum since those VMFs represents similar sound units.
iii) The VMFs in each group are summed and NLM estimation is performed to remove the noise components. The grouping of VMFs reduces the computational cost.
iv) Finally, the NLM estimated signals obtained from each of the groups are combined to obtain the enhanced signal.

This method depends upon the NLM estimation of the VMFs. In the following pages, a brief introduction to VMD and a discussion on the need for grouping of VMFs is presented. Then, NLM estimation for removing noise components from VMFs is discussed.

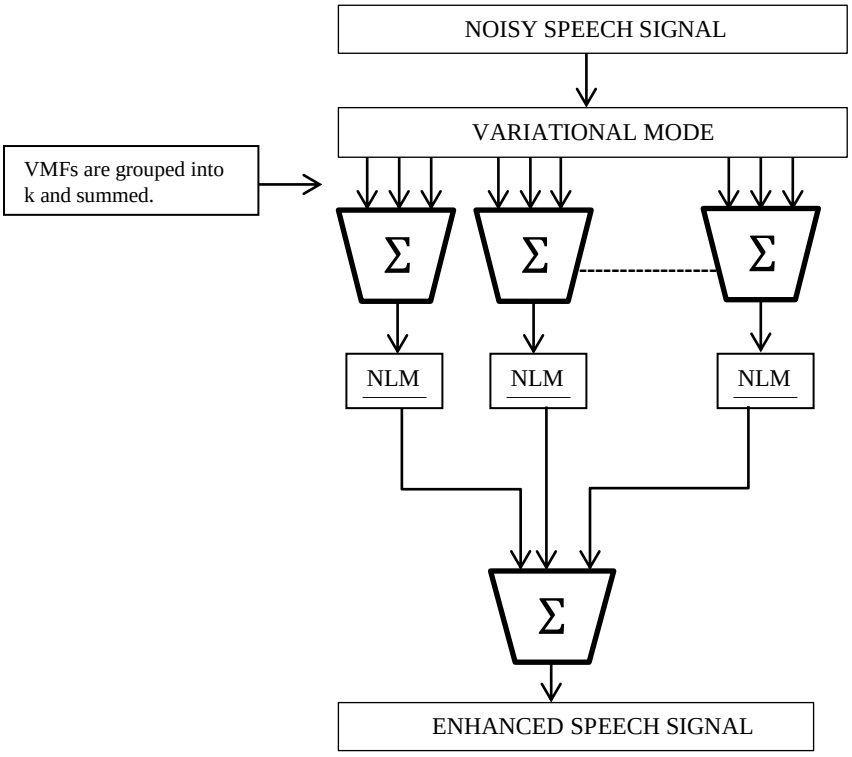

Figure 1: The above block diagram represents the proposed method for enhancing the speech signal.

- The block diagram shows that the noisy signal is getting decomposed into a set of Variational Mode Function's (namely: - VMF_1, VMF_2, VMF_3 …VMF_K).
- The VMF's having lower center frequency represents the high magnitude regions and the VMF's with higher center frequency represents the low magnitude regions.
- The VMF's are divided into set of j groups based on the similarity in their center frequencies. Now they are summed and NLM estimation is performed in order to remove the noisiness in the speech signal. The groups are finally combined to obtain the enhanced speech signal.

2.1. VARIATIONAL MODE DECOMPOSITION

In this section we introduce our proposed model for variational mode decomposition, essentially based on the three concepts outlined in the previous section. The goal of VMD is to decompose a real valued input signal f into a discrete number of sub-signals (modes) that have specific sparsity properties while reproducing the input. Here, the sparsity prior of each mode is chosen to be its bandwidth in spectral domain.

In other words, we assume each mode k to be mostly compact around a center pulsation ωk, which is to be determined along with the decomposition. In order to assess the bandwidth of a mode, we propose the following scheme**:**

1) For each mode ωk, compute the associated analytic signal by means of the Hilbert transform in order to obtain a unilateral frequency spectrum.

2) For each mode, shift the mode's frequency spectrum to "baseband", by mixing with an exponential tuned to the respective estimated center frequency.

3) The bandwidth is now estimated through the H1 Gaussian smoothness of the demodulated signal, i.e. the squared L2-norm of the gradient. The resulting constrained variational problem is the following:

$$\min_{\{u_k\},\{\omega_k\}} \left\{ \sum_k \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) * u_k(t) \right] e^{-j\omega_k t} \right\|_2^2 \right\} \quad \text{s.t.} \quad \sum_k u_k = f$$

Eq(1)

Such that Pk vk(t) = s(t). Where, {vk} = {v1, v2 ...vk}, {ωk} = {ω1, ω2 ...ωk}, k, δ (t) and $*$ represents the VMFs (modes), the center frequencies for each of the VMFs, total number of modes, Dirac distribution and convolution operator, respectively.

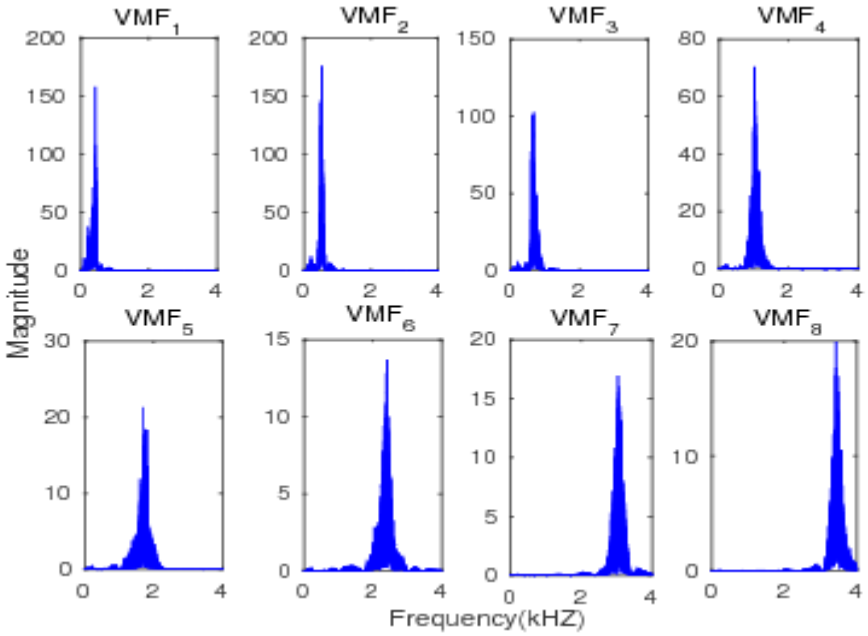

Figure 2: Magnitude spectrum of VMFs for a 0 dB white noise added speech signal. The modes are arranged from low- to high frequency band (left to right).

Variational method for decomposing a signal into an ensemble of band-limited intrinsic mode functions, that we call Variational Mode Decomposition, (VMD). In contrast to existing decomposition models, like the empirical mode decomposition (EMD), we refrain from modeling the individual modes as signals with explicit IMFs. Instead, we replace the most recent definition of IMFs, namely their characteristic description as AM-FM signals, by the corresponding narrow-band property.

If a large number of modes are selected for decomposition, under-binning of modes (loss of information) happens. On the other hand, lower number of modes results in over-binning of modes (mode duplication). During the preliminary experiments performed on development set, it was observed that for effective decomposition and reconstruction of speech signal, a minimum of k = 12 levels of decomposition is required. The magnitude spectra for the 12 VMFs derived from a 0dB white noise added speech signal are shown in Figure 2. The magnitude spectra shown from left to right in ascending order of VMFs.

It can be observed that, in the each of the VMFs, frequency and amplitude variations are very small. It can also be noted that, depending upon the similarities in the location of their center frequency and mean magnitude, some of the VMFs can be combined together.

For example, VMF−1 & VMF−2 can be combined to represent a single group. The VMFs are combined to reduce the computational cost for NLM estimation without loss in denoising capability. In this study, the VMFs are finally clustered into four groups.

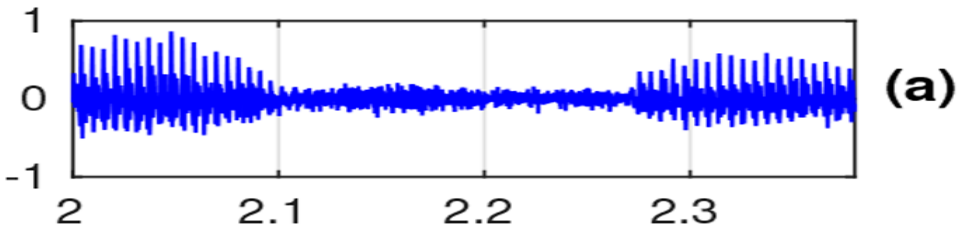

Figure 3: represents the segment of speech taken with 0db white noise.

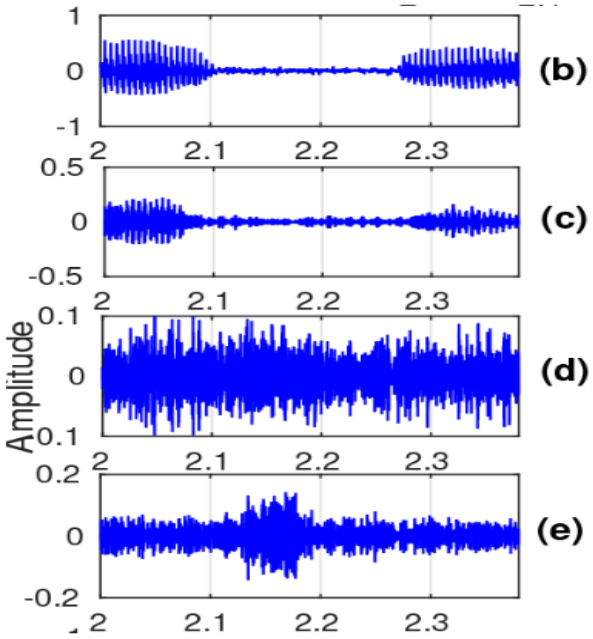

Figure 4: from (b-e) represents the four groups of VMF's obtained by combining original VMF's.

2.2 NON-LOCAL MEANS ESTIMATION

The NLM approach estimates the true signal from the noisy signal by exploiting the non-l similarities among the sample points. In NLM filtering, for each sample point of the signal x(n) estimate xˆ(n) is computed as a weighted sum of the signal values at another sample point x(m). final denoised signal is computed with the help of two local patches with starting points being n m, respectively. Both the patches consist of P samples and they lie within the search neighborh N(n). The estimated denoised signal is computed.

$$\hat{x}(n) = \frac{1}{W(n)} \sum_{m \epsilon N(n)} w(n,m)x(m)$$

Eq(2)

For each sample point, the mapping is decided by weight values w(n, m) that represent the non-l similarity present in the neighborhood with respect to the sample points x(n) and x(m), respectiv The weight value w(n, m)is computed.

$$\omega(n,m) = \exp\left(-\frac{\sum_{j=1}^{P}(s(n+j)-s(m+j))^2}{2PB^2}\right)$$

Eq(3)

where, B represents the bandwidth parameter which controls the amount of smoothing to be app to the denoised signal. The difference values are summed over P samples (length of the patch) normalized in order to get the weight value. W(n) represents the normalized weight value at sam point n which, in turn, is computed as follows:

$$W(n) = \sum_{m \epsilon N(n)} w(n,m)$$

Eq(4)

By following the above process of enhancement the noise speech signal i.e., Fig (3) and Fig(4) is denoised as shown in the below figures.

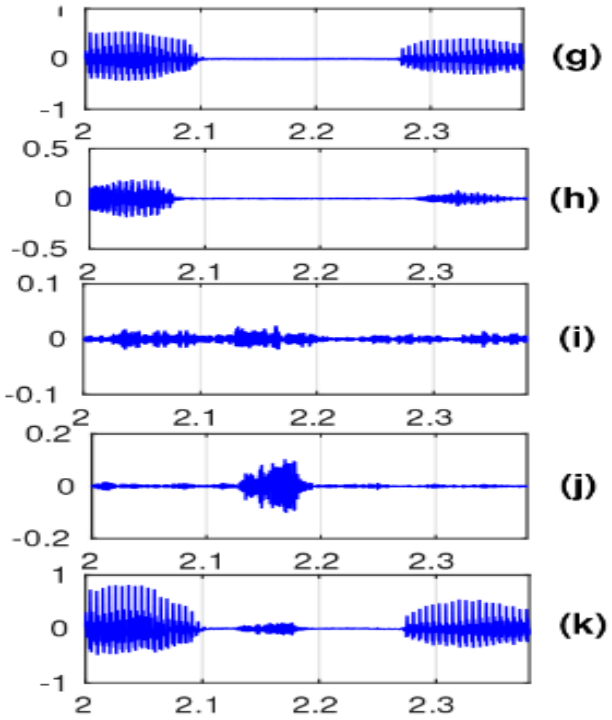

Figure 5: from (g)-(j) VMFs after denoising using NLM estimation, (f) the original clean signal (k) enhanced signal obtained by proposed approach.

2.3 Final speech enhancement by NLM estimation of VMF's

In the case of speech, the amplitude and the frequency change over the frames depending on the sound units. Therefore, the NLM is not effective in enhancing noisy speech signal. those variations are suppressed.

to a great extent by grouping the VMFs. The NLM estimation is performed on the signal obtained by adding the VMFs belonging to any particular group. The final reconstruction is done by adding each of the NLM estimated outputs as shown in Figure(1).

The effectiveness of the proposed approach for speech enhancement is demonstrated in Figure (3), (4) & (5). It is evident that, the fluctuations in each group of VMFs are very less. The NLM effectively removes the noise components from the VMFs. By comparing the original clean and enhanced speech signals, it is evident that the proposed approach is very effective in removing the noise components from the given speech data. The below figure showing the NLM estimation of VMF.

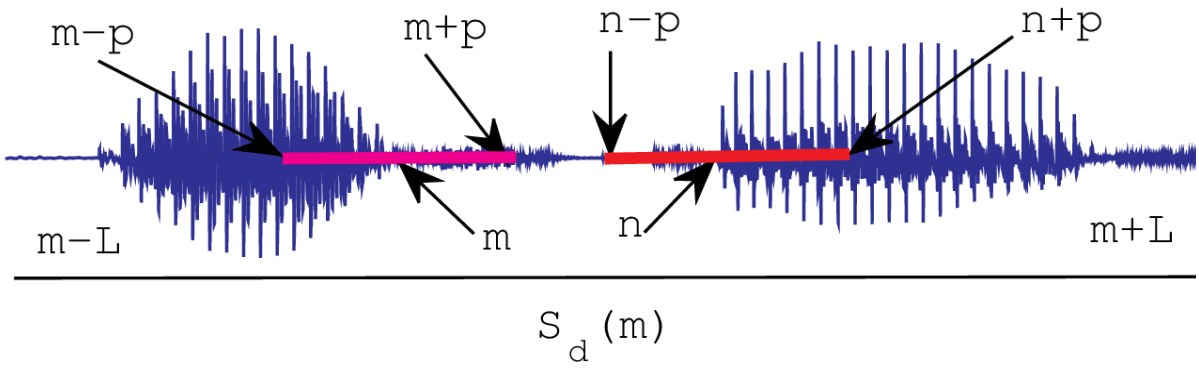

Figure 6: NLM estimation & construction of patches from neighborhood having non-local similarities.

3. EXPERIMENTAL RESULTS

We have applied 8-level decomposition of noisy speech signal using VMD technique. For VMD, the data fidelity constraint balancing parameter was set 320; time-step was 0 while tolerance of convergence was selected as 10^−7. The NLM estimation is dependent on proper selection of some tunable parameters like patch size (P), search neighborhood size N(n), and bandwidth parameter (B). In this study, the value of P, N(n)and B are selected as 10, 200 and 0.4σ, respectively on first group VMFs. Similarly P, N(n) and B are selected as 10, 100 and 0.6σ on second group. For third and fourth groups those parameters are selected as 20, 80 and 0.8σ, respectively. Where σ represents the standard deviation of the summed signal of respective group of VMFs. All the tunable parameter values were selected empirically.

Noise	SNR in dB	segSNR		PESQ	
		FBE	VMD-NLM	FBE	VMD-NLM
White	10	4.58	5.77	2.56	3.01
	5	3.09	4.80	2.34	2.85
	0	2.18	3.87	2.03	2.51

Table 1

In order to evaluate the efficacy of the existing and proposed approaches, speech signals from the TIMIT database were used. A set of 10 speech utterances from 5 male and 5 female speakers was used for experimental evaluations. The clean speech files were corrupted by adding white noise at three different levels of signal to noise ratios (0, 5, and 10dB). These non-stationary background noise sources were obtained from the Noisex-92 database.

The following objective speech quality measures were used for evaluating the performance: perceptual evaluation of speech quality (PESQ) and segmental signal to noise ratio (segSNR). The results of the experimental evaluations are given in Table 1. Compared to the existing approaches, the proposed speech enhancement technique is expected to result in better segSNR and PESQ values especially for low SNR values (i.e., 0 and 5 dB). Consistent improvements are noted for all the three noise types explored in this study. The best case performances are presented in boldface to highlight the same. Expect for 10dB white noise case, the proposed approach is expected to be significantly better.

4. CONCLUSION AND FUTURE SCOPE

In this report, a two-stage VMD-NLM based speech enhancement technique has been explored. The noisy speech signal is first decomposed into 8 VMFs using the VMD algorithm. Next, based on the similarities in the location of center frequencies and the mean amplitudes, the VMFs are clustered and summed to yield a set of four VMFs. This step reduces the overall computational cost. The so obtained VMFs are then processed through NLM estimation in order to effectively reduce the ill-effects of interfering noises. This approach is compared with two of the recently developed speech enhancement techniques in terms of objective speech quality measures like segSNR and PESQ. The noise applied at different SNR levels are used for experimental evaluation. The proposed speech enhancement approach is expected to be better than the explored methods.

Adaptive algorithms found effective for such cases and are utilized in this problem. The work is based on decomposition method using variational mode decomposition (VMD) technique, where the decomposed components signify the frequency characteristics of the signal. Since Wiener filtering is used in VMD inherently, it can be modified with the least mean squares (LMS) adaptive algorithm for good accuracy and adaptability in this work. Different noises like Babble noise, Street noise, and Exhibition noise can be considered and the corresponding signals can be decomposed into set of intrinsic mode functions (IMFs). Basically, the lower modes are of high frequency and noisy; whereas the higher mode IMFs contain the low and medium frequency components and are considered as the enhanced signal. The results of the proposed algorithm are found excellent as compared to earlier techniques. The resultant wave forms are visually observed and the sound is verified for audible range. Also different measuring parameters are considered for its performance measure.

It is measured in terms of signal-to-noise ratio (SNR), segmental signal to noise ratio (SegSNR), perceptual evaluation of speech quality (PESQ) and log spectral distance (LSD). The technique is verified with standard database NOIZEUS for 0, 5, 10dB respectively and also in real world case.

Also, we are looking forward to work with other techniques such as adaptive variational mode

decomposition which is a bit different from the decomposition method used in this work, and EMD Wavelet method of denoising signals which includes the basic idea that the energy of a signal will often be concentrated in a few coefficients in wavelet domain while the energy of noise is spread among all coefficients in wavelet domain as the hard and soft thresholding methods for denoising, where the former leaves the magnitudes of coefficients unchanged if they are larger than a given threshold, while the latter just shrinks them to zero by the threshold value.

6. REFERENCES

[1] Nagapuri Srinivas, Gayadhar Pradhan and S Shahnawazuddin, “Enhancement of Noisy Speech Signal by Non-Local Means Estimation of Variational Mode Functions.” Interspeech 2018, 2-6 September 2018, Hyderabad, 2018.

[2] P. C. Loizou, Speech enhancement: theory and practice. CRC press, 2013.

[3] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, “An overview of noise-robust automatic speech recognition,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 4, pp. 745–777, 2014.

[4] J. Ming, T. J. Hazen, J. R. Glass, and D. A. Reynolds, “Robust speaker recognition in noisy conditions,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 5, pp. 1711–1723, 2007.

[5] S. 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.

[6] Y. Lu and P. C. Loizou, “Estimators of the magnitude-squared spectrum and methods for incorporating snr uncertainty,” IEEE transactions on audio, speech, and language processing, vol. 19, no. 5, pp. 1123–1137, 2011.

[7] 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.

[8] R.Tavares and R.Coelho,“Speech enhancement with non-stationary acoustic noise detection in time domain,” IEEE Signal Processing Letters, vol. 23, no. 1, pp. 6–10, 2016.

[9] B.Yegnanarayana, C.Avendano, H.Hermansky, and P.S.Murthy, “Speech enhancement using linear prediction residual,” Speech Communication, vol. 28, no. 1, pp. 25–42, may 1999. [14] 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, 1999.

[10] K. Deepak and S. M. Prasanna, “Foreground speech segmentation and enhancement using glottal closure instants and mel cepstral coefficients,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 7, pp. 1205–1219, 2016.

[11] N. Chatlani and J. J. Soraghan, “EMD-based filtering (EMDF)lio of low-frequency noise for speech enhancement,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 4, pp. 1158–1166, 2012.

[12]K.Khaldi,A.-O.Boudraa,andA.Komaty,“Speech enhancement using empirical mode decomposition and the teager–kaiserenergy operator,” The Journal of the Acoustical Society of America, vol. 135, no. 1, pp. 451–459, 2014

.

.