Interspeech 2018
2-6 September 2018, Hyderabad

Enhancement of Noisy Speech Signal by Non-Local Means Estimation of Variational Mode Functions

Nagapuri Srinivas, Gayadhar Pradhan and S Shahnawazuddin

Department of Electronics and Communication Engineering
National Institute of Technology Patna, India.
(ns, gdp, s.syed)@nitp.ac.in

Abstract

In this paper, a speech enhancement approach exploiting the efficacy of non-local means (NLM) estimation and variational mode decomposition (VMD) is proposed. 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) 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 two existing speech enhancement techniques. The experimental results presented in this study show that, the proposed method provides better speech enhancement performance.

Index Terms: Speech enhancement, noisy speech, non-local means, variational mode function.

1. Introduction

With the recent development of machine learning algorithms, the primary focus of research in speech processing is to create robust human-machine interactive systems. The speech signal used for the development of automatic speech and speaker recognition systems, in most of the cases, is degraded by ambient noises present in the recording environment and communication channel [1]. The performance of those systems reduce significantly when the test data is noisy [2,3]. Therefore, speech enhancement is an essential component for developing robust speech-based user applications. 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 [4–6]. In those approaches, short-time noise spectrum is estimated from the non-speech regions determined using voice activity detection (VAD) module. Then, the estimate of the noise spectrum is subtracted from the noisy speech spectrum to enhance the signal quality [4–6]. The performance of such approaches is highly dependent on the accuracy with which the non-speech region are detected and robust estimation of instantaneous noise spectrum [7, 8]. Several techniques have been proposed for estimating the noise spectrum from the noisy speech signal [9–11]. However, such spectral enhancement methods introduce distortion in the enhanced speech signal due to deviations in estimated and actual instantaneous noise spectrum [8, 12]. In the enhancement approaches presented in [13–16], the high signal to noise ratio (SNR) regions are identified and relatively more enhanced compared to the low SNR regions. The linear prediction (LP) residual signal corresponding to the small regions around the instants of significant excitation are weighted to enhance those regions relative to other portions. The speech signal is reconstructed using the modified LP residual signal. Such temporal enhancement methods are not efficient in completely removing the background noise from the noise degraded speech signals [16].

Recently, several adaptive signal decomposition methods like empirical mode decomposition (EMD) and it's variants have been proposed for suppressing stationary and non-stationary noises from the noisy speech signal [17–20]. The combination of EMD and variational mode decomposition (VMD) has also been explored for speech enhancement [21]. This method is effectively reduce the low-frequency noise as well as high-frequency noise. However, those signal decomposition methods are not effective when the speech signal is corrupted by speech-like noises [21].

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 [22, 23]. 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 paper by utilizing the efficacy of VMD and NLM estimation.

The remainder of this paper is organized as follows: The proposed method for speech enhancement using NLM estimation of VMFs is presented in Section 2. The experimental studies for evaluating the performance of the proposed and existing techniques are presented in Section 3. Finally, the paper is concluded in Section 4.

10.21437/Interspeech.2018-1928

2. Proposed speech enhancement approach

The block diagram summarizing the proposed method for speech enhancement is shown in Fig 1. In the proposed 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 represent 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.

The method proposed in this study primarily depends upon the NLM estimation of the VMFs. In the following sub-sections, 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.

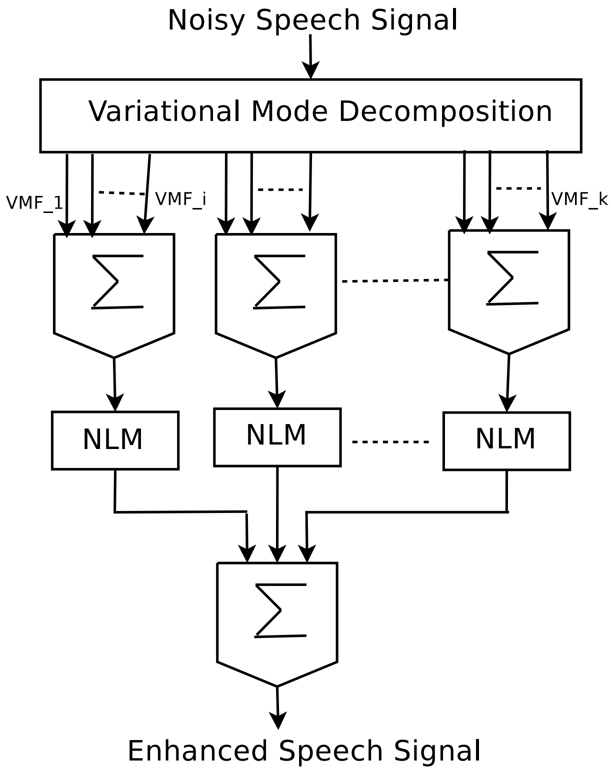

Figure 1: *The block diagram representing proposed method for enhancing speech signal.*

2.1. Variational mode decomposition of noisy speech

The VMD is a non-recursive, concurrent signal decomposition method that breaks the given input signal $(s(t))$ into several modes termed as VMFs [24]. Each VMFs (v_k) represents a narrow-band frequency region of the input signal. The VMD also estimates the center frequency (ω_k) of each VMFs as H^1-norm. The center frequencies are sparsity priors which helps in reconstruction of input signal $s(t)$. The v_k and ω_k are computed by solving the constrained variational problem as follows:

$$\min_{\{v_k\},\{\omega_k\}} \left\{ \sum_k \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) * v_k(t) \right] e^{-j\omega_k t} \right\|_2^2 \right\} \quad (1)$$

such that $\sum_k v_k(t) = s(t)$. Where, $\{v_k\} = \{v_1, v_2, ...v_k\}$, $\{\omega_k\} = \{\omega_1, \omega_2, ...\omega_k\}$, k, $\delta(t)$ and $*$ represents the VMFs (modes), the center frequencies for each of the VMFs, total number of modes, Dirac distribution and convolution operator, respectively.

The signal reconstruction constraint is addressed by using Lagrangian multipliers (λ) and the quadratic penalty factor (α). The convergence properties of the penalty term at a finite weight value and strict enforcement of constraint by the Lagrangian multiplier are being utilized. The augmented Lagrangian $\mathcal{L}$ is represented as follows:

$$\mathcal{L}(\{v_k\}, \{\omega_k\}, \lambda) = \alpha \sum_k \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) * v_k(t) \right] e^{-j\omega_k t} \right\|_2^2 + \left\| s(t) - \sum_k v_k(t) \right\|_2^2 + \left\langle \lambda(t), s(t) - \sum_k v_k(t) \right\rangle \quad (2)$$

By using augmented Lagrangian and the alternate direction method of multipliers optimization framework, the VMFs and corresponding center frequencies can be computed. After optimization, the resultant updated modes $\{\hat{v}_k\}$ in frequency domain are computed as follows:

$$\hat{v}_k^{n+1}(\omega) = \frac{\hat{s}(\omega) - \sum_{i \neq k} \hat{v}_i(\omega) + \frac{\hat{\lambda}(\omega)}{2}}{1 + 2\alpha(\omega - \omega_k)^2} \quad (3)$$

where $\hat{v}(w)$, $\hat{s}(w)$ and $\hat{\lambda}(w)$ are the frequency domain representations of $v_k(t)$, $s(t)$ and $\lambda(t)$, respectively. The modes in time domain, $v_k(t)$ can be obtained from $\hat{v}_k(\omega)$ using the inverse Fourier transform. Similarly, the updated center frequencies are optimized in Fourier domain as follows:

$$\omega_k^{n+1} = \frac{\int_0^\infty \omega |\hat{v}_k(\omega)|^2 d\omega}{\int_0^\infty |\hat{v}_k(\omega)|^2 d\omega} \quad (4)$$

It locates the updated frequency which is at the center of the k^{th} mode power spectrum.

2.2. Grouping VMFs to reduce variations

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) [24]. 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

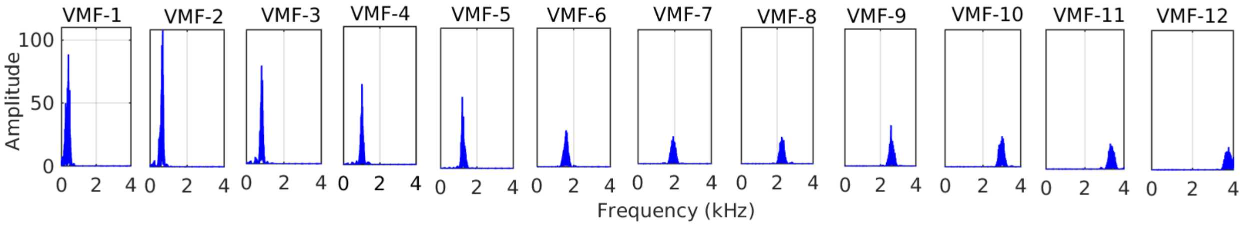

Figure 2: *Magnitude spectrum of VMFs for a* 0 *dB white noise added speech signal. The modes are arranged from low- to high-frequency band (left to right).*

their center frequency and mean magnitude, some of the VMFs can be combined together. For example, $VMF{-}3$ to $VMF{-}5$ 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.

2.3. NLM estimation

The NLM approach estimates the true signal from the noisy signal by exploiting the non-local similarities among the sample points. In NLM filtering, for each sample point of the signal $x(n)$, an estimate $\hat{x}(n)$ is computed as a weighted sum of the signal values at another sample point $x(m)$. The final denoised signal is computed with the help of two local patches with starting points being n and m, respectively. Both the patches consist of P samples and they lie within the search-neighborhood $N(n)$. The estimated denoised signal is computed as follows [25]:

$$\hat{x}(n) = \frac{1}{W(n)} \sum_{m \epsilon N(n)} w(n,m)x(m) \tag{5}$$

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

$$\omega(n,m) = \exp\left(-\frac{\sum_{j=1}^{P}(s(n+j)-s(m+j))^2}{2PB^2}\right) \tag{6}$$

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

$$W(n) = \sum_{m \epsilon N(n)} w(n,m) \tag{7}$$

2.4. Final speech enhancement by NLM estimation of VMFs

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. However, as discussed in Section 2.2, 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. It is evident that, the fluctuations in each group of VMFs is 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. Similar inferences can be drawn by comparing the spectrograms for clean, noisy and enhanced speech signals shown in Figure 4.

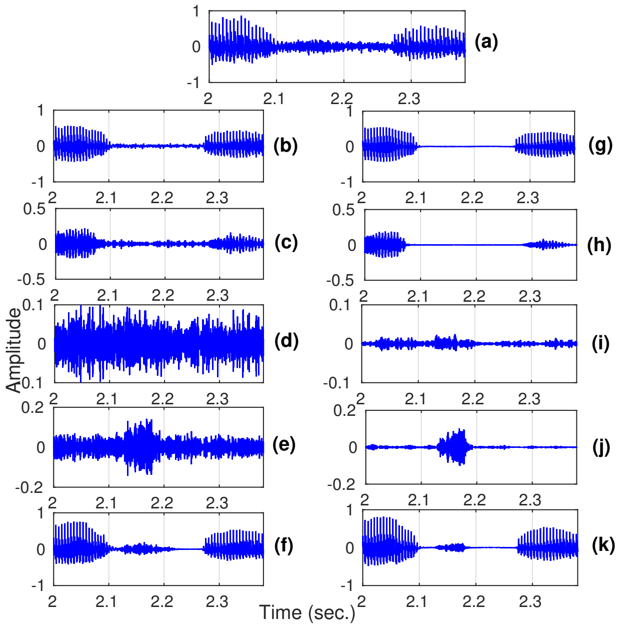

Figure 3: *The plots illustrate enhancement of noisy speech signal by using propose method. (a) A segment of speech taken from TIMIT database with* 0 *dB white noise added to it. (b)-(e) the four groups of VMFs obtained by combining original VMFs. (g)-(j) VMFs after denoising using NLM estimation, (f) the original clean signal (k) enhanced signal obtained by proposed approach.*

3. Results and discussions

We have applied 12-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 pa-

Table 1: *Performance evaluation of the proposed and existing speech enhancement techniques in terms of scale of background intrusiveness (BAK), scale of the mean opinion score (OVL), segmental signal to noise ratio (segSNR) and perceptual evaluation of speech quality (PESQ). The performances are evaluated after degrading the speech data with white, factory and babble noises. For each cases, three different SNR values are chosen.*

Noise	**SNR**	**BAK**			**OVL**			**segSNR**			**PESQ**		
	in dB	**FBE**	**EMD-VMD**	**Prop.**	**FBE**	**EMD-VMD**	**Prop.**	**FBE**	**EMD-VMD**	**Prop.**	**FBE**	**EMD-VMD**	**Prop.**
White	**10**	2.57	3.23	**3.30**	3.05	3.19	**3.28**	4.58	**5.81**	5.77	2.56	2.71	**3.01**
	5	2.36	2.70	**2.96**	2.73	2.85	**3.01**	3.09	4.59	**4.80**	2.34	2.40	**2.85**
	0	2.07	2.23	**2.76**	2.33	2.14	**2.74**	2.18	2.66	**3.87**	2.03	2.19	**2.51**
Factory	**10**	2.37	2.73	**2.92**	2.82	2.75	**2.83**	4.30	4.34	**5.41**	2.45	2.57	**2.71**
	5	2.12	2.18	**2.54**	2.40	2.39	**2.49**	2.97	2.52	**3.38**	2.23	2.35	**2.46**
	0	1.83	1.68	**1.93**	2.16	2.05	**2.27**	-0.62	-0.77	**-0.22**	2.02	1.98	**2.29**
Babble	**10**	2.21	2.79	**2.70**	2.55	2.61	**2.73**	4.54	4.42	**5.26**	2.36	**2.44**	2.41
	5	1.93	2.24	**2.31**	2.19	2.52	**2.86**	2.64	2.66	**3.08**	2.17	2.02	**2.14**
	0	1.61	1.72	**1.78**	1.81	1.96	**2.15**	-0.86	-1.01	**-0.30**	1.79	1.85	**1.92**

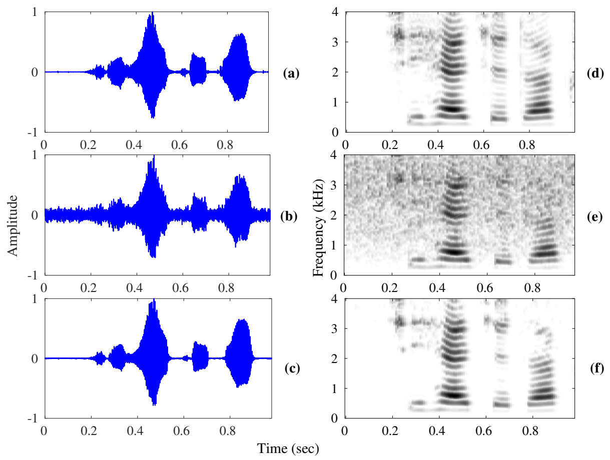

Figure 4: *(a) A segment of clean speech signal taken from TIMIT database. (b) The signal after adding 0 dB white noise. (c) Enhanced signal obtained by using the proposed method. (e)-(f) Spectrograms for clean, noisy and enhanced speech signals, respectively.*

rameters 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.

The proposed approach is compared with two existing speech enhancement techniques reported in [16, 21]. The enhancement technique reported in [16], is motivated by the fact that, the characteristics of the interfering sources vary with respect to time. Consequently, the interfering background noise can temporally overlap with the desired speech or it can exists as an isolated event in the recorded signal. To address this issue, a two stage approach was proposed in that work. Fist the foreground speech was segmented from rest of the background noise. Then, the LP analysis was performed on foreground speech. The regions around the glottal closure instants in the LP residual signal and the LP formants were then modified to reconstruct the enhanced speech. In rest of the paper this method is termed as FBE. In [21], an effective combination of VMD and EMD techniques was explored for speech enhancement. EMD was used to break the noisy speech signal into a number of intrinsic mode functions (IMFs). Next, a set of IMFs were summed up and VMD was then applied on summation of selected IMFs. This speech enhancement method is referred to as EMD-VMD in this paper.

In order to evaluate the efficacy of the existing and proposed approaches, speech signals from the TIMIT database [26] 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, factory noise and babble noise at three different levels of signal to noise ratios (0, 5, and 10 dB). These non-stationary background noise sources were obtained from the Noisex-92 database [27]. The following objective speech quality measures were used for evaluating the performance: perceptual evaluation of speech quality (PESQ) [28], scale of background intrusiveness (BAK) [28], scale of the mean opinion score (OVL) [28] and segmental signal to noise ratio (segSNR) [29].

The results of the experimental evaluations are given in Table 1. Compared to the existing approaches, the proposed speech enhancement technique is noted to result in better BAK, OVL, 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 and 10dB babble noise cases, the proposed approach is observed to be significantly better.

4. Conclusion

In this paper, a two-stage VMD-NLM based speech enhancement technique has been proposed. The noisy speech signal is first decomposed into 12 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. The proposed approach is compared with two of the recently developed speech enhancement techniques in terms of objective speech quality measures like BAK, OVL, segSNR and PESQ. Three different noise types at different SNR levels are used for experimental evaluation.The proposed speech enhancement approach is observed to be better than the explored methods.

5. References

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

[2] 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.

[3] 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.

[4] 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.

[5] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. ICASSP*, vol. 4, 1979, pp. 208–211.

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

[7] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Transactions on speech and audio processing*, vol. 11, no. 5, pp. 466–475, 2003.

[8] 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.

[9] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE transactions on acoustics, speech, and signal processing*, vol. 33, no. 2, pp. 443–445, 1985.

[10] 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.

[11] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, 2012.

[12] R. Tavares and R. Coelho, "Speech enhancement with nonstationary acoustic noise detection in time domain," *IEEE Signal Processing Letters*, vol. 23, no. 1, pp. 6–10, 2016.

[13] 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.

[15] P. Krishnamoorthy and S. M. Prasanna, "Enhancement of noisy speech by temporal and spectral processing," *Speech Communication*, vol. 53, no. 2, pp. 154–174, 2011.

[16] 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.

[17] 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.

[18] K. Khaldi, A.-O. Boudraa, and A. Komaty, "Speech enhancement using empirical mode decomposition and the teager–kaiser energy operator," *The Journal of the Acoustical Society of America*, vol. 135, no. 1, pp. 451–459, 2014.

[19] L. Zao, R. Coelho, and P. Flandrin, "Speech enhancement with emd and hurst-based mode selection," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 5, pp. 899–911, 2014.

[20] K. Khaldi, A.-O. Boudraa, A. Bouchikhi, and M. T.-H. Alouane, "Speech enhancement via EMD," *EURASIP Journal on Advances in Signal Processing*, vol. 2008, no. 1, p. 873204, 2008.

[21] A. Upadhyay and R. Pachori, "Speech enhancement based on mEMD-VMD method," *Electronics Letters*, vol. 53, no. 7, pp. 502–504, 2017.

[22] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *Proc. CVPR*, vol. 2, 2005, pp. 60–65.

[23] P. Singh, G. Pradhan, and S. Shahnawazuddin, "Denoising of ECG signal by non-local estimation of approximation coefficients in DWTchat," *Biocybernetics and Biomedical Engineering*, vol. 37, no. 3, pp. 599–610, 2017.

[24] K. Dragomiretskiy and D. Zosso, "Variational mode decomposition," *IEEE transactions on signal processing*, vol. 62, no. 3, pp. 531–544, 2014.

[25] B. H. Tracey and E. L. Miller, "Nonlocal means denoising of ecg signals," *IEEE transactions on biomedical engineering*, vol. 59, no. 9, pp. 2383–2386, 2012.

[26] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, 1993.

[27] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[28] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on audio, speech, and language processing*, vol. 16, no. 1, pp. 229–238, 2008.

[29] ——, "Evaluation of objective measures for speech enhancement," in *Ninth International Conference on Spoken Language Processing*, 2006.