Enhancing Noisy Speech using WEMD

J.V. Thomas Abraham, A. Shahina, A. Nayeemulla Khan

Abstract*: Speech signal distortion is unavoidable in real time applications. This distorted signal can adversely affect the performance of systems based on speech signals. Automatic speaker recognition (ASR) system performs well with clean speech signals while its performance degrades drastically with noisy speech. Enhancing the speech signal aims at improving the quality of the speech signal by reducing the noise contamination, thereby improving the performance of the ASR system. Noise could be background noise, reverberation, babble noise etc. In this paper, to improve the distorted speech signal, we propose a two stage speech enhancement algorithm where Empirical Mode Decomposition (EMD) with adaptive threshold in IMF selection is done at the first stage and then employ wavelet denoising (WD) in the second stage. The two stage denoising method is used to reduce noise in high and low frequencies. The effectiveness of the proposed algorithm is compared with a few baseline algorithms used for enhancement.

***Index Terms*: Speech enhancement, Empirical mode decomposition, wavelet denoising,**

I. INTRODUCTION

Speech signals in real time applications get distorted by several factors. Background noises, the reverberation of the signal, or people speaking in surrounding may all corrupt the actual speech signal and thereby affect the performance of the applications that use these signals. The motive behind any speech enhancement algorithm is to reduce the noise level effectively such that it results in a cleaner signal, without any loss of information contained in it. The need for denoising a speech signal depends on the purpose for which it is being used namely:

- To improve the speech signal so that automatic speech or speaker recognition systems performs better
- To improve the signal quality so that it is more acceptable to human listeners.
- To improve the understandability of the speech signal.
- To improve the speech signals such that it can be encoded more effectively for storage or transmission.

Many enhancement techniques were proposed that includes minimum mean square error (MMSE) estimation [1], spectral subtraction [2], subspace methods [3] and Wiener filter [4]. These techniques generally assume that the background noise is stationary, and they perform better for high signal-to-noise (SNR) values. The simplest method for signal enhancement is to adopt some type of linear filters to remove the noise from the signal. However the drawback with this technique is that the noise spectrum should be known a-priori to construct the appropriate filter and the signal should be stationary. Both these conditions are not applicable in case of real signals. The problem arises if the noise is non-stationary, and the signal has low signal-to-noise ratio (SNR) values, and most of the real world signals are nonlinear and non-stationary. Both WD and EMD are powerful tools for analyzing nonlinear and non-stationary signals like speech [5]. Recently, deep learning based speech enhancement methods have achieved state-of-art results in this research field [6].

This paper is structured as follows: Section II gives a brief overview of EMD decomposition, Intrinsic Mode Function (IMF) and WD. Section III explains the proposed enhancement method for the noisy speech signal. Section IV explains the speech enhancement experiments to evaluate the proposed approach. Finally, Section V concludes the work and discusses the extension of the proposed work.

II. OVERVIEW OF WD AND EMD

WD eliminates a considerable amount of noise while preserving the important features in the signal. Wavelet denoising involves appropriate selection of the wavelet function or main wavelet, filter and then choosing the level of signal decomposition [7]. The Wavelet denoising gives time variant decomposition of the signal, hence, depending on the characteristics of the different event related responses we can choose different wavelet coefficients for different time ranges [8]. Although, wavelets achieve satisfactory time-frequency resolution, their effectiveness depends on the basis function to be specified and the choice of the basis functions is not always an easy one.

Huang of NASA proposed a new method called Empirical Mode Decomposition for signal processing, especially for non-linear and non-stationary signals [9]. This approach is completely data driven and doesn't need any a-priori knowledge of the noise for signal decomposition. EMD decomposition is fully adaptive and is based on the local properties of the signal. Since Speech signals are non-stationary and non-linear, EMD works better on these signals and also reconstruction of signals using IMFs is complete.

The EMD will split the signal into IMFs and each IMF is a function that (i) has exactly one extreme between zero crossings and (ii) has a zero mean. To obtain the IMFs, sifting process is applied on the signal. The sifting process is explained below:

Find the mean m1 of lower and upper envelop of the speech signal s(t). We used cubic-spline interpolation of local maxima and minima.

- Determine the difference d1 between the signal and the mean m1. ie., d1 = s(t) - m1
- In the next iteration of the sifting process, let d_1 be the signal, and m_{11} is the mean of d_1's upper and lower envelopes. Calculate d_{11} as d_1-m_{11}
- Repeat the sifting process k times, until d_{1k} is an IMF, that is $d_{1(k-1)}$-m_{1k} = d_{1k}

Revised Version Manuscript Received on April 05, 2019.

J.V. Thomas Abraham, School of Computing Science and Engineering, VIT University Chennai Campus, Chennai, India.
A. Shahina, Department of Information Technology**,** SSN College of Engineering, Kalavakkam, Chennai, India.
A. Nayeemulla Khan, School of Computing Science and Engineering, VIT University Chennai Campus, Chennai, India.

- Let $c_1 = d_{1k}$, be the signal's first IMF. The first IMF contains the shortest period component of the signal.
- Remove the first IMF, c_1, from rest of the signal: $r_1 = s(t)-c_1$. Repeat this procedure on r_j: $r_2 = r_1-c_2$,... $r_n = r_{n-1}-c_n$

The sifting process stated above results in a set of functions called IMFs and the number of IMFs depends on the input signal. The IMFs got by the above sifting process have the characteristic such that the initial IMFs have high frequency and gradually decrease to the last, hence the noise characterized by high frequency is mainly found on the starting IMFs and reduces towards later IMFs. So with this method, noise can be reduced by filtering or thresholding some IMFs [10].

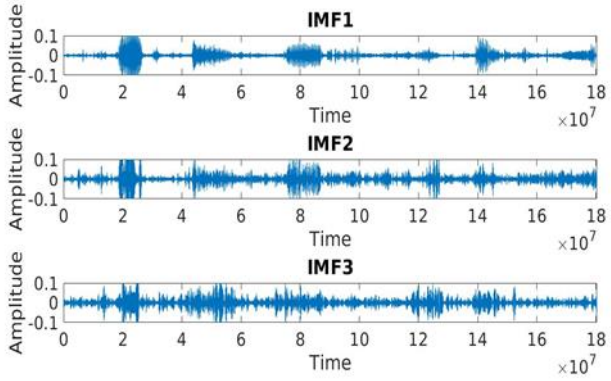

Fig 1: IMFs (1-3) of a Speech Signal distorted with Babble Noise at 0dB SNR

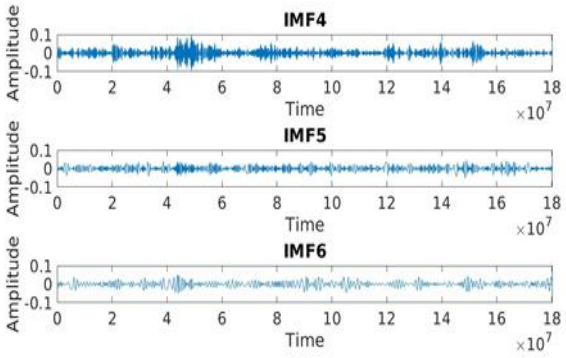

Fig 2: IMFs (4-6) of a Speech Signal distorted with Babble Noise at 0dB SNR

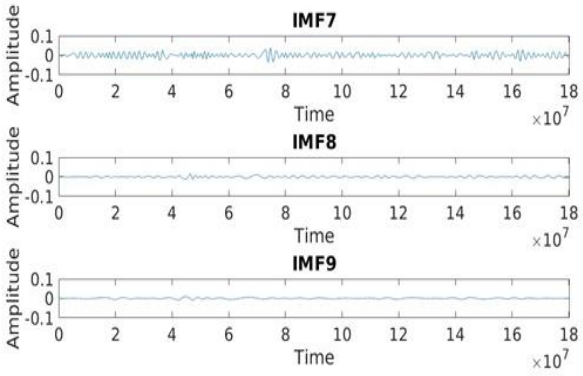

Fig 3: IMFs (7-9) of a Speech Signal distorted with Babble Noise at 0dB SNR

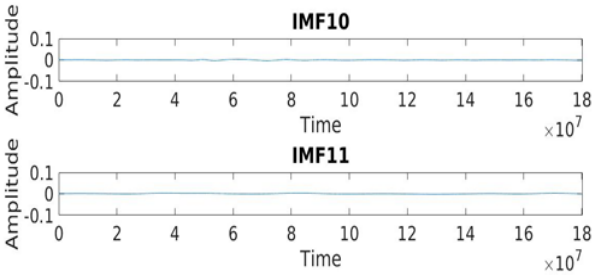

Fig 4: IMFs (10-11) of a Speech Signal distorted with Babble Noise at 0dB SNR

From Figure 1 we can observe faster oscillations in the first IMF, slightly slower oscillations in the second IMF and slower fluctuations in the third, and so on. That is, at each time interval, the EMD algorithm separates between high-frequency IMFs and low-frequency IMFs. Thus, the high-frequency signal exists in the first IMF. In other words, the lower order IMFs indicate faster oscillations (high frequency modes), and higher order IMFs indicate slower oscillations (low frequency modes) in EMD [11].

III. PROPOSED ALGORITHM (WEMD)

In this letter we take the advantage of wavelet denoising and the EMDH algorithm and propose a two stage speech enhancement method. In the first stage, we apply EMD on the noisy speech signal and decompose it into a set of IMFs. Let us assume that we get *n* IMFs. Then we select the IMF based on the Hurst exponent calculated from the frames, within each IMF. This short-time segment analysis avoids any sudden changes in the non-stationary noises' power spectrum which could affect the IMF selection of the entire speech signal. Then, add all the IMFs and get the signal *es(t)*. Now the output signal *es(t)* is enhanced further using wavelet denoising.

The wavelet denoising allows us to have a distributed representation of several real-world signals. That is the wavelet denoising concentrates on large-magnitude wavelet coefficients which are generally signals. The small wavelet coefficients typically characterize noise and we dwindle those coefficients or eliminate them without affecting the signal. After we threshold the coefficients, the enhanced signal is reconstructed using the inverse wavelet transform. In our experiment we use empirical Bayesian method with a Cauchy prior as the denoising method and the 'sym4' wavelet is used with a posterior median threshold rule. The block diagram of the proposed algorithm is shown in Fig 5.

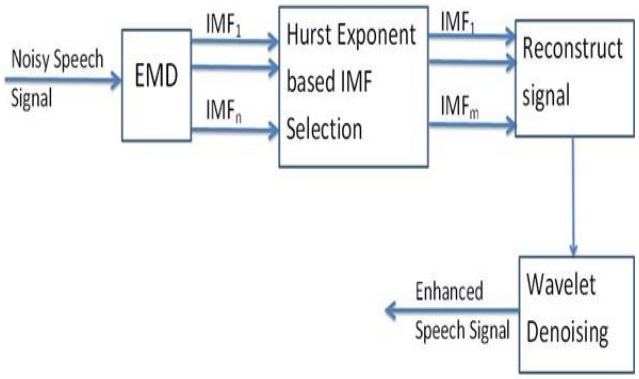

Fig 5: Block diagram of the WEMD Algorithm

WEMD Algorithm

Step 1: Decompose the noisy speech signal *s(t)* into intrinisic mode functions using EMD IMF_i, i = 1 .. n
Step 2: For each IMF find the hurst exponent and discard IMF with low-frequency noise components.
Step 3: Reconstruct the signal *es(t)*, by adding all IMFs.
Step 4: The stage one enhanced signal *es(t)* is decomposed into low and high frequency components, resulting in wavelet coefficients.
Step 5: Select the small value coefficients and shrink or remove them without affecting the signal.
Step 6: Perform inverse wavelet transform and reconstruct the signal to get the final enhanced signal, *ees(t)*.

IV. Experimental Setup and Results

The clean TIMIT speech signals are corrupted with additive background noises (babble, airport and street) at various signal-to-noise ratios (0, 5, 10, 15) to get a noisy speech signal. The background noises from NOIZEUS dataset and FreeSound.org is used. The WEMD algorithm is then applied on the noisy speech signal to get the enhanced speech signal. We consider the performance measures like Short Time Objective Intelligibility (STOI) (lower is better), Segmental SNR (preferrbly higher), Mean Square Error (MSE) (should be low) and PESQ (higher the better), to compare the efficiency of the proposed algorithm.

Table 1: For Babble Noise

Method	SNR(in)	STOI	SegSNR	MSE	PESQ
WEMD	0	86.985118	-4.533045	0.001573	1.90
EMDH	0	86.913876	-4.649207	0.001710	1.93
WEMD	5	97.303032	-1.889909	0.000502	2.20
EMDH	5	97.437007	-1.986746	0.000544	2.19
WEMD	10	99.225847	1.030321	0.000162	2.49
EMDH	10	99.319643	0.949705	0.000175	2.46
WEMD	15	99.69644	4.183401	0.000055	2.77
EMDH	15	99.738874	4.123258	0.000059	2.72

Table 2: For Airport Noise

Method	SNR(in)	STOI	SegSNR	MSE	PESQ
WEMD	0	91.643885	-4.584223	0.001582	2.04
EMDH	0	93.110128	-4.642340	0.001660	2.03
WEMD	5	98.052678	-1.951696	0.000502	2.27
EMDH	5	98.453743	-2.006549	0.000528	2.24
WEMD	10	99.433875	1.041624	0.000161	2.54
EMDH	10	99.539738	0.986650	0.000170	2.50
WEMD	15	99.785570	4.237136	0.000054	2.81
EMDH	15	99.814532	4.186584	0.000057	2.76

Table 3: For Street Noise

Method	SNR(in)	STOI	SegSNR	MSE	PESQ
WEMD	0	95.353204	-4.491481	0.001679	1.96
EMDH	0	95.694534	-4.540716	0.001742	1.96
WEMD	5	98.575091	-1.833443	0.000534	2.31
EMDH	5	98.818491	-1.876877	0.000555	2.29
WEMD	10	99.461839	1.088140	0.000172	2.60
EMDH	10	99.548669	1.052748	0.000179	2.56
WEMD	15	99.745041	4.250545	0.000059	2.88
EMDH	15	99.780050	4.231978	0.000061	2.81

Tables 1-3 show the average value of the specified measures for eight speakers (sp01 to sp08) for the babble, airport and street noises. It is seen for all the performance measures the proposed method is better than the existing method. Table 4 also shows the average value of SNR(out) we see the proposed method performs better than the two other existing methods EMDH and mEMD-VMD [12].

Table 4: Average SNR(out) of Enhanced Speech Signal

	Babble Noise			
Method / SNR (in)	**0**	**5**	**10**	**15**
EMDH	0.185	4.366	9.302	14.184
mEMD-VMD	0.387	4.661	9.683	14.432
WEMD	0.460	5.420	10.330	15.020
	Airport Noise			
Method / SNR (in)	**0**	**5**	**10**	**15**
EMDH	0.194	4.363	9.313	14.047
mEMD-VMD	0.398	4.660	9.692	14.527
WEMD	0.440	5.420	10.350	15.080
	Street Noise			
Method / SNR (in)	**0**	**5**	**10**	**15**
EMDH	0.010	4.350	9.301	14.071
mEMD-VMD	0.317	4.826	9.688	14.383
WEMD	0.170	5.140	10.050	14.730

V. CONCLUSION

The enhancement algorithm proposed in this letter uses fusion of wavelet denoising and empirical mode decomposition. The IMFs resulting from EMD is selected using Hurst exponent. The proposed method works for both low and high SNR signals and for many types of real world noises. The WEMD algorithm was tested with babble, airport and street noises with SNR ranging from 0dB to 15dB. The measures like STOI, Segmental SNR, SNR of enhanced signal, mean square error, and PESQ calculated for the proposed algorithm produced better results than the existing algorithm.

REFERENCES

1. K. K. Paliwal, B. Schwerin, and K. K. Wójcicki, "Speech enhancement using a minimum mean-square error short-time spectral modulation magnitude estimator," Speech Communication, vol. 54, pp. 282–305, 2012.
2. S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," Acoustics, Speech and Signal Processing, IEEE Transactions on, vol. 27, pp. 113–120, 05 1979.
3. K. Hermus, P. Wambacq, and H. Van hamme, "A review of signal subspace speech enhancement and its application to noise robust speech recognition," EURASIP J. Appl. Signal Process., vol. 2007, no. 1, pp. 195–195, 2007.
4. M. A. Abd El-Fattah, M. I. Dessouky, A. M. Abbas, S. M. Diab, E.-S. M. El-Rabaie, W. Al-Nuaimy, S. A. Alshebeili, and F. E. Abd El-Samie, "Speech enhancement with an adaptive wiener filter," Int. J. Speech Technol., vol. 17, no. 1, pp. 53–64, 2014.
5. R. Tavares and R. Coelho, "Speech enhancement with non-stationary acoustic noise detection in time domain," IEEE Signal Processing Letters, vol. 23, pp. 6–10, 2016.
6. Y. Xu, J. Du, L. Dai, and C. Lee, "A regression approach to speech enhancement based on deep neural networks," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, pp. 7–19, Jan 2015.
7. V. S. Cherkassky and S. Kilts, "Myopotential denoising of ECG signals using wavelet thresholding methods," Neural Networks, vol. 14, pp. 1129–1137, 2001.
8. D.-F. Guo, W.-H. Zhu, Z.-M. Gao, and J.-Q. Zhang, "A study of wavelet thresholding denoising," in WCC 2000 - ICSP 2000. 2000 5th International Conference on Signal Processing Proceedings. 16th World Computer Congress 2000, vol. 1, pp. 329–332, 2000.
9. N. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, 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," Proceedings of the Royal Society of London A: mathematical, physical and engineering sciences, vol. 454, pp. 903–995, 1998.
10. N. Chatlani and J. Soraghan, "Emd-based filtering (EMDF) of low-frequency noise for speech enhancement," IEEE Transactions on Audio, Speech and Language Processing, vol. 20(4), pp. 1158–1166, 2012.
11. L. ZÃčo, R. Coelho, and P. Flandrin, "Speech enhancement with EMD and hurst-based mode selection," Audio, Speech, and Language Processing, IEEE/ACM Transactions on, vol. 22, pp. 899–911, 2014.
12. A. Upadhyay and R. B. Pachori, "Speech enhancement based on mEMD-VMD method," Electronics Letters, vol. 53, no. 7, pp. 502–504, 2017.