EMPIRICAL MODE DECOMPOSITION FOR JOINT DENOISING AND DEREVERBERATION

Tariqullah Jan and Wenwu Wang

Centre for Vision, Speech and Signal Processing,
University of Surrey, GU2 7XH, UK
(t.jan@surrey.ac.uk, w.wang@surrey.ac.uk)

ABSTRACT

We propose a novel algorithm for the enhancement of noisy reverberant speech using empirical-mode-decomposition (EMD) based subband processing. The proposed algorithm is a one-microphone multistage algorithm. In the first step, noisy reverberant speech is decomposed adaptively into oscillatory components called intrinsic mode functions (IMFs) via an EMD algorithm. Denoising is then applied to selected high frequency IMFs using EMD-based minimum mean-squared error (MMSE) filter, followed by spectral subtraction of the resulting denoised high-frequency IMFs and low-frequency IMFs. Finally, the enhanced speech signal is reconstructed from the processed IMFs. The method was motivated by our observation that the noise and reverberations are disproportionally distributed across the IMF components. Therefore, different levels of suppression can be applied to the additive noise and reverberation in each IMF. This leads to an improved enhancement performance as shown in comparison to a related recent approach, based on the measurements by the signal-to-noise ratio (SNR).

1. INTRODUCTION

Room reverberation is one of the main causes of performance degradation in automatic speech recognition (ASR) systems. It is commonly modeled as the combination of three parts, the direct signal, early reflections and the late reflections [1, 4]. The direct signal is the main signal from the speaker to the microphone. The early reflections deteriorate the speech spectrum due to the nonflat frequency response, while late reflections degrade the quality and intelligibility of speech. Late reverberation can cause serious problems to ASR performance.

The late reverberations are usually treated as noise whose variance is estimated and then subtracted from the reverberant speech, for which the spectral subtraction (SS) technique has been widely used [1]. To estimate the late reverberations, a method based on an exponential decay function has been developed in [12]. The main challenge in suppression of late reverberations is to estimate accurately its variance. The presence of noise from the acoustical environments make it more difficult to estimate the power of late reverberations. Therefore, in this paper, we consider to enhance the noisy reverberant speech by jointly dealing with the late reverberations and the additive acoustic noise having a Gaussian distribution and white spectrum.

We propose a new method using EMD based subband analysis. We use an EMD algorithm to decompose the noisy reverberant speech into a linear combination of the so-called intrinsic mode functions (IMFs), ranging from the high-frequency to low-frequency bands [2], [3]. Then we select the IMFs that have higher levels of noise and apply the EMD based MMSE filter [8] to reduce the additive noise. In the next step, we use the denoised IMF components and the remaining IMF components to estimate the power of late reverberations. We have observed that the energy of the late reverberations is spread over the different IMFs with different magnitude. For this reason, we perform spectral subtraction to each IMF according to the energy of the late reverberations present in the IMF components. We will show the improved enhancement performance with the proposed method. The next section presents our proposed approach in detail. Section 3 shows the evaluation results, followed by a conclusion in Section 4.

2. SYSTEM DESCRIPTION

Our proposed dereverberation system is depicted in Figure 1. First, the EMD algorithm [3] is applied to the noisy reverberant speech $x(t)$ to decompose the signal adaptively into C IMF components $\tilde{z}_j(t)$. In the next step R components are selected from the C IMF components of $\tilde{z}_j(t)$ for denoising. Then, an EMD based MMSE filter [8] is applied to each of the selected IMFs to reduce its noise level. Spectral subtraction with variable scaling factors is applied to the denoised IMFs and the remaining IMFs separately. Finally the signal is reconstructed as $\hat{s}(t)$.

2.1 EMD analysis

The EMD algorithm describes the signal details at certain frequency bands in the form of different IMFs [5]. Each IMF has a distinct time scale and acts as a basis function [3]. There are two main conditions that need to be satisfied by each IMF [3]. First, the difference between the number of extrema and the number of zero crossings should not exceed one. Second, the average value for the envelope assigned to the local maxima and minima is zero. We perform subband analysis of the noisy reverberant signal $x(t)$ using the following EMD algorithm [3]:

1) Determine all extrema of $x(t)$.

2) Compute the "envelopes" of the maxima and minima as $\alpha_{min}(t)$ and $\alpha_{max}(t)$ by interpolation.

3) Find the average, $r_C(t) = (\alpha_{min}(t) + \alpha_{max}(t))/2$.

4) Extract the detail $\tilde{z}(t) = x(t) - r_C(t)$.

5) Repeat steps 1-4 for the residue $r_C(t) = x(t) - \tilde{z}(t)$.

A *sifting* process is applied to refine the above procedure corresponding to the steps 1-4 until $\tilde{z}(t)$ can be considered as zero mean according to some stopping criterion [3]. Once this is achieved, $\tilde{z}(t)$ can be considered as an effective IMF. Finally the residue $r_C(t)$ is computed and step 5 is applied.

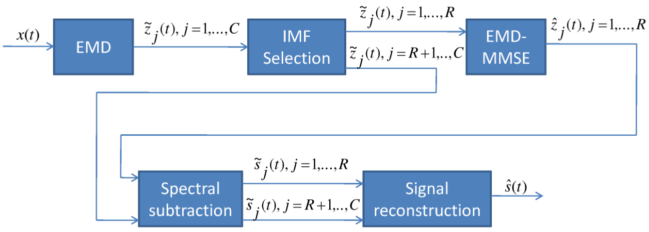
$x(t)$
EMD
$\tilde{z}_j(t), j=1,...,C$
IMF Selection
$\tilde{z}_j(t), j=1,...,R$
$\tilde{z}_j(t), j=R+1,..,C$
EMD-MMSE
$\hat{z}_j(t), j=1,...,R$
Spectral subtraction
$\tilde{s}_j(t), j=1,...,R$
$\tilde{s}_j(t), j=R+1,..,C$
Signal reconstruction
$\hat{s}(t)$

Figure 1: Block diagram of our proposed denoising and dereverberation system.

Upon convergence of the algorithm, $x(t)$ is decomposed into a sum of C IMFs and a residue $r_C(t)$,

$$x(t) = \sum_{j=1}^{C} \tilde{z}_j(t) + r_C(t) \tag{1}$$

where $\tilde{z}_j(t)$ represents the jth IMF component. Typically, C was set to 15 in our simulations, where different values of C have also been tested which however give similar results.

2.2 IMF selection

We use the selected IMFs for the denoising in the next subsection 2.3. In order to explain the reason behind the selection of these IMFs, we present an example in which we first generate the noisy speech signal by adding white Gaussian noise to the clean speech signal at SNR= 4 dB. Then the EMD algorithm is used to derive the IMF components of the above generated signal. In Figure 2 the first three high frequency IMFs and the last three low frequency IMFs of this signal are shown. From this figure, it can be observed that the noise is mainly present in the high frequency IMF components. Motivated from this observation we choose the high frequency IMF components $\tilde{z}_j(t), j = 1,...,R$ for denoising. In our experiments we used R=10, which is found empirically to be a suitable number.

2.3 EMD-MMSE

In this step, we perform denoising for the selected high frequency IMF components $\tilde{z}_j(t)$, where $j = 1,...,R$, using the MMSE filter [8]. In general speech noise can be estimated using Boll's method [10]. The silence periods of the signal are detected and then the noise power spectrum is estimated by averaging the power spectra of the noisy signal on the M first temporal frames corresponding to the silence period. Here we used the first R IMFs separately in order to estimate the noise power, following the relation [8]

$$\mid \hat{B}_j(k) \mid^2 = \frac{1}{M} \sum_{i=0}^{M-1} \mid B_j(k;i) \mid^2, \qquad j = 1,...,R \tag{2}$$

where $\mid B_j(k;i) \mid$ represents the magnitude spectrum of the jth IMF component at the discrete frequency k and time frame i (index used for the silence period), and $\mid \hat{B}_j(k) \mid^2$ is the estimated noise power of the jth IMF component at frequency bin k.

The combined operation of EMD and MMSE filter [7, 9] is named as EMD-MMSE. Hence each IMF is filtered by the MMSE filter as follows:

$$\hat{z}_j(k;n) = H_j(k;n)\tilde{z}_j(k;n), \qquad j = 1,...,R \tag{3}$$

where $\hat{z}_j(k;n)$ and $\tilde{z}_j(k;n)$ are the spectra of the jth estimated IMF and noisy IMF components respectively, observed at the discrete frequency k and the time frame n. $H_j(k;n)$ can be defined as follows [7]

$$H_j(k;n) = \frac{SNR_{prio}(k;n)}{1 + SNR_{prio}(k;n)} \tag{4}$$

The signal to noise ratio, SNR_{prio} can be estimated based on the previous frame of the estimated $\hat{z}_j(k;n-1)$ and a local estimation of SNR_{inst}, given as [7]

$$\begin{aligned} &SNR_{prio}(k;n) \\ &= \alpha \frac{\hat{z}_j^2(k;n-1)}{\hat{B}_j{}^2(k)} + (1-\alpha) max(SNR_{inst}(k;n), 0) \end{aligned} \tag{5}$$

where α is a weighting factor (chosen empirically to be 0.98 in this work) and SNR_{inst} represents the instantaneous SNR, and can be defined as the local estimation of SNR_{prio},

$$SNR_{inst} = \frac{\tilde{z}_j^2(k;n)}{\hat{B}_j{}^2(k)} \tag{6}$$

Hence $\hat{z}_j(k;n)$ with $j = 1,...,R$, obtained in equation (3) are the denoised IMF components which are further processed in the next step in order to remove the late reverberations from these components.

2.4 Spectral subtraction for the IMFs

It has been observed that the late reverberations lead to the blurring effect on speech spectrum in the frequency domain, resulting in a smoothed spectrum [1]. Therefore, the power spectrum of the late reverberation components can be estimated as the smoothed and shifted version of the power spectrum of the denoised reverberant IMF components $\hat{z}_j(k,n), j = 1,...,R$ and remaining low frequency components, $\tilde{z}_j(k,n), j = R+1,...,C$. For simplicity, all of these components are now represented by $\hat{z}_j(k,n)$ where $j = 1,...,C$.

$$|S_{l_j}(k;n)|^2 = \gamma\omega(n-\rho) * |\hat{z}_j(k;n)|^2 \tag{7}$$

where $|S_{l_j}(k;n)|^2$ is the short term power spectrum of the late reverberations in the jth IMF component, γ is the scaling factor specifing the relative strength of the late reverberation components, the symbol $*$ denotes the convolution operation, $\omega(n)$ is a smoothing window, and ρ refers to the relative delay of the late reverberations. The short-term speech spectrum can be obtained by using the Hamming window of length 16 msec with 8 msec overlap for the short-term Fourier analysis.

To estimate the power spectrum of the original speech, we can subtract the power spectrum of the late reverberation components from that of the IMF components $\hat{z}_j$, $j = 1,...,R$. Spectral subtraction can be employed for each selected component as follows [1],

$$\begin{aligned} &|\tilde{s}_j(k;n)|^2 = \\ &|\hat{z}_j(k;n)|^2 max\left[\frac{|\hat{z}_j(k;n)|^2 - \gamma_j\omega(n-\rho) * |\hat{z}_j(k;n)|^2}{|\hat{z}_j(k;n)|^2}, \varepsilon\right] \end{aligned} \tag{8}$$

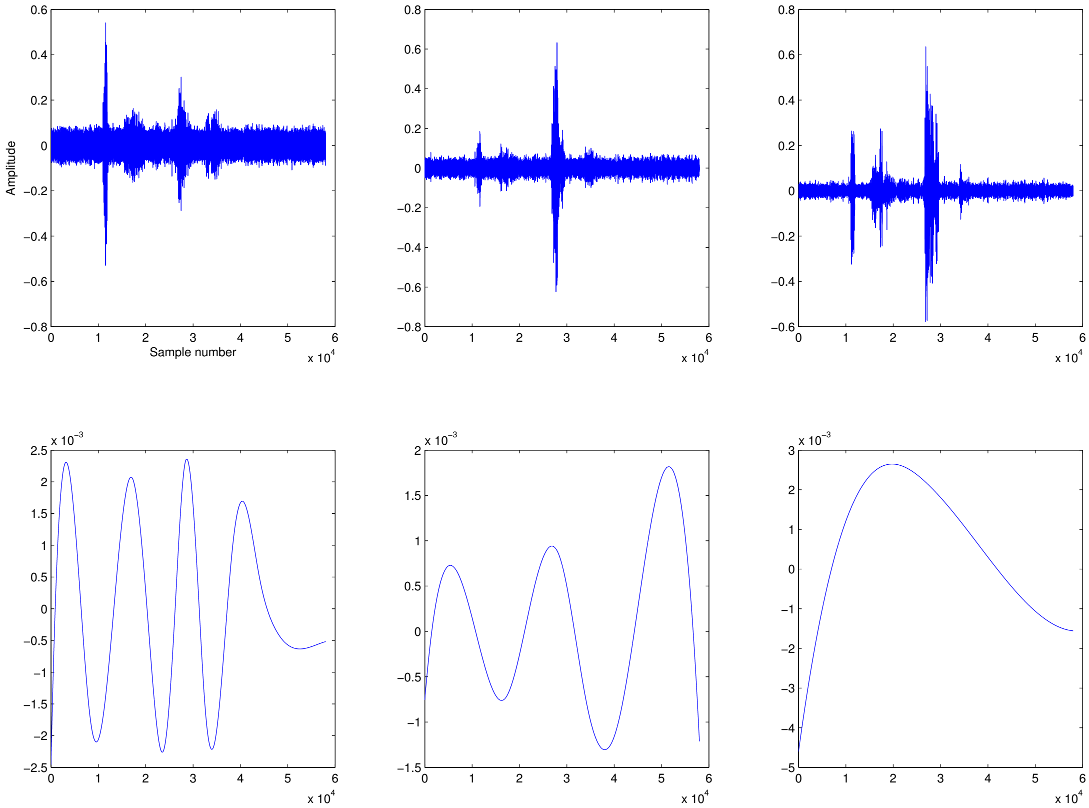

Figure 2: The IMF components derived from the noisy speech signal. The top row shows the first three high frequency IMFs and the bottom row shows the last three low frequency IMFs.

where $|\tilde{s}_j(k;n)|^2$ represents the power spectrum of the jth IMF component of the estimated version of the original speech, ε stands for the floor parameter which was set to be 0.001 in our experiments, corresponding to the maximum attenuation of 30 dB and γ_j is a scaling factor, discussed below. The spectral subtraction procedure discussed above in equation (7) and (8) were also used for all the IMF components including the remaining low frequency IMFs $\tilde{z}_j, \quad j = R+1, ..., C$.

2.5 Selection of variable scaling factor γ_j

We use the variable scaling factor γ_j for the estimation of the late reverberations from the IMF components. To show the motivation for using variable γ_j, we present an example in which we take the IMF components of the reverberant speech signal (at RT= 200 msec) and the clean speech signal. We then subtract the IMF components of clean speech signal from the corresponding IMF components of the reverberant signal to obtain the distribution of the energy of late reverberations. The spectrograms of the subtracted IMF components are shown in Figure 3. From this figure, it can be observed that the late reverberations tend to spread over the different IMFs with variable energy, i.e. having high energy in the first few high frequency IMFs and decreases in the lower IMFs. Motivated by this fact, we propose to use variable scaling factors γ_j instead of a fixed one (as used in method [1]). We select high values of γ for the first few high frequency IMF components while decreasing them for the lower frequency components. We have tested different range of values for γ. The optimized ranges of values of γ for each corresponding IMF component are shown in Figure 4 where reverberation time (RT) is equal to 200 and 500 msec respectively.

2.6 Signal reconstruction

Finally, the enhanced signal $\hat{s}(t)$ can be reconstructed by the superposition of the processed IMFs, and the residue, given as follows,

$$\hat{s}(t) = \sum_{j=1}^{R} \tilde{s}_j(t) + \sum_{j=R+1}^{C} \tilde{s}_j(t) + r_C(t) \tag{9}$$

where $\tilde{s}_j(t)$ is computed as the inverse FFT of $\tilde{s}_j(k;n)$ obtained in (8).

3. EXPERIMENTAL RESULTS AND DISCUSSIONS

In this section, we evaluate the performance of our proposed method using simulations. Four clean speech utterances, 2 male and 2 female all sampled at 16 kHz were used. The simulated room model [6] was used to generate the reverberant signals from the clean speech signals with different RTs,

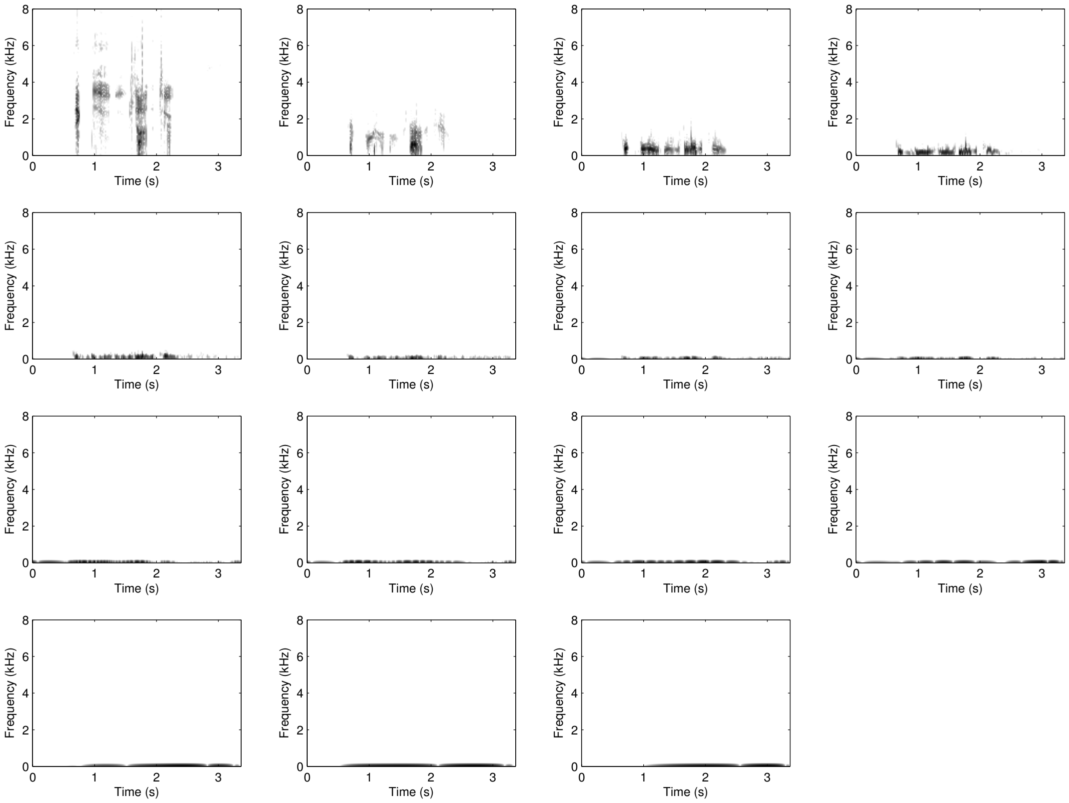

Figure 3: The spectrograms of the subtracted IMFs shown in the descending order of frequency patterns with the highest frequency component on the top left and the lowest frequency component on the bottom right.

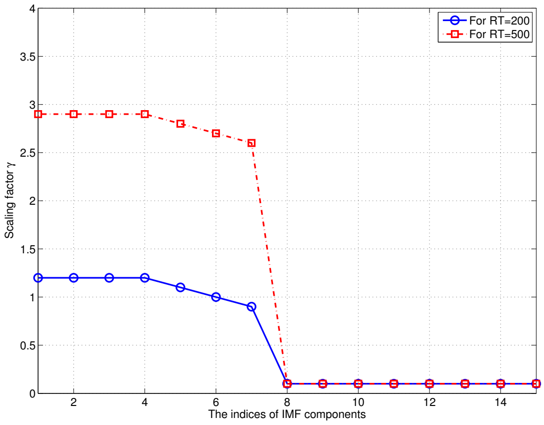

Figure 4: Variable scaling factor γ_j, j=1,...,15

which were then added by white Gaussian noise with SNR values ranging from -12 dB to 4 dB. The size of the room in the experiments was $10 \times 10 \times 10$, and the microphone and speaker were positioned at [3, 8, 5] and [2, 2, 5] respectively (the unit is meter) [6]. The performance index used in the evaluations is the SNR [11]. The SNR can be defined as,

$$SNR = 10log_{10}\frac{\sum_{i=1}^{T}(s(t_i))^2}{\sum_{i=1}^{T}(s(t_i)-\hat{s}(t_i))^2} \qquad (10)$$

where $s(t_i)$ and $\hat{s}(t_i)$ are the original signal and the enhanced signal respectively, and T is the length of the signal.

We have performed numerical simulations for RT= 200 and 500 msec respectively, with SNR ranging from -12 dB to 4 dB for each RT. In total 50 independent random tests have been conducted for each SNR, and the average results were calculated. In order to ensure the fair comparison between our proposed approach and the method in [1], EMD-MMSE has also been applied as a preprocessing step for the method in [1]. Figure 5 shows the comparison of the methods for the signals in terms of SNR obtained for RT= 200 and 500 msec respectively, and for different noise levels. From Figure 5, we can observe that our proposed algorithm offers improvement over the method in [1] with EMD-MMSE preprocessing, especially for RT equal to 500 msec, and comparable performance is observed for RT equal to 200 msec.

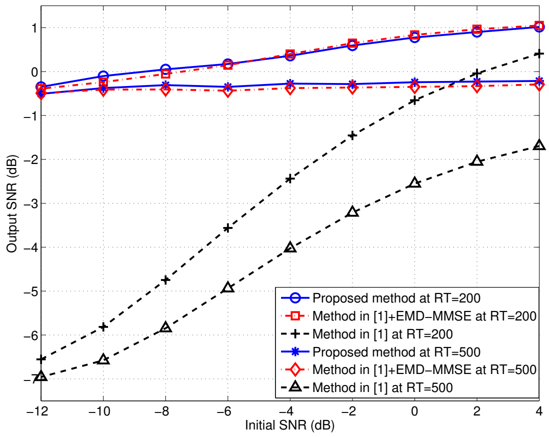

Figure 5: Average gain in SNR for *RT*= 200 msec and 500 msec with different initial noise levels. Results are the average of 50 random tests.

As compared to the results obtained by [1] without incorporating EMD-MMSE preprocessing, our proposed method has shown considerably higher performance improvement.

We performed another set of experiments in which we evaluate and compare the performance of the proposed approach and the method in [1] with and without EMD-MMSE filtering on the basis of different source-microphone distances. The *RT* used in this set of experiments for all the four signals is 500 msec with initial *SNR*= -4 dB. Average results for all the speech signals based on 50 random tests, are depicted in Figure 6. We can observe that as the distance between the source and the microphone decreases the average performance of both algorithms increases. In addition, it should be noted that our proposed method performs better for larger source-microphone distances.

4. CONCLUSION

A novel approach has been presented for speech denoising and dereverberation, based on the EMD decomposition of the noisy reverberant speech. EMD based MMSE and spectral subtraction have been used for processing the IMF components separately. It has been observed that both the additive noise and the late reverberations are spread over the different IMF components in varying magnitudes. As shown in our experiments, performing spectral subtraction for each of these components offers better denoising and dereverberation performance as compared with a related method that directly uses the noisy reverberant speech.

Acknowledgment

T. U. Jan was supported by the NWFP UET Peshawar Pakistan, and W. Wang was supported in part by the EPSRC of the UK. Both authors are grateful to M. Wu and D.L. Wang of [1], and Z. Wu and N.E. Huang of [2] for making their matlab codes publicly available.

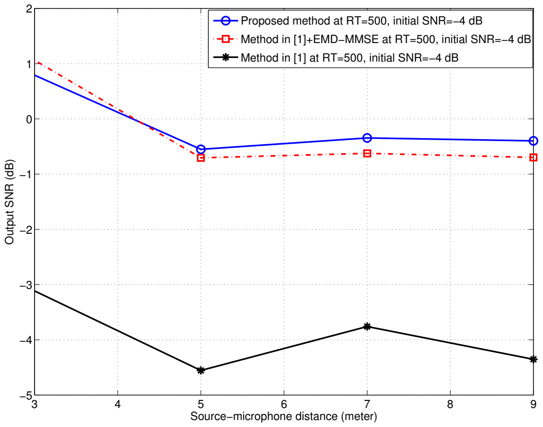

Figure 6: Average gain in SNR for different source-microphone distances where *RT*= 500 msec with initial noise level equal to -4 dB. Results are the average of 50 random tests.

REFERENCES

[1] M. Wu and D. L. Wang, “A two-stage algorithm for one-microphone reverberant speech enhancement,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, pp. 774–784, May 2006.

[2] Z. Wu and N. E. Huang, “Ensemble empirical mode decomposition: a noise assisted data anlysis method,” *Advances in Adaptive Data Analysis*, vol. 1, pp. 1–41, Jul. 2008.

[3] N. E. Huang, Z. Shen, S. R. Long, et al., “The empirical mode decomposition and Hilbert spectrum for nonlinear and non-stationary time series analysis,” in *Proceedings of the Royal Society A*, vol. 454, pp. 903–995, 1998.

[4] B. Yegnanarayana and P. S. Murthy, “Enhancement of reverberant speech using LP residual signal,” *IEEE Trans. Speech Audio Process.*, vol. 8, no.3, pp. 267–281, May 2000.

[5] P. Flandrin, G. Rilling, and P. Goncalves, “Empirical mode decomposition as a filter bank,” *IEEE Signal Process. Letters*, vol. 11, no. 2, pp. 112–114, 2004.

[6] J. B. Allen and D. A. Berkley “Image method for efficiently simulating small-room acoustics,” *J. Acoust. Soc. Amer.*, vol. 65, 1979.

[7] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error short time spectral amplitude estimator,” *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. 32, no. 6, pp. 1109–1121, 1984.

[8] K. Khaldi, A. O. Boudraa, A. Bouchikhi and M. T. Alouane, “Speech enhancement via EMD,” *EURASIP J. Adv. Signal Process.*, vol. 2008, Mar 2008.

[9] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error Log-spectral amplitude estimator,” *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. ASSP-33, no. 2, 1985.

[10] S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” in *IEEE Trans. Acoust., Speech, and Signal Process.*,vol. 27, no. 2, pp. 113–120, 1979.

[11] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, “Objective Measures of Speech Quality,”, Prentice Hall, Englewood Cliffs, NJ, USA, 1988,

[12] B. Kingsbury and N. Morgan, “Recognizing reverberant speech with rasta-plp,” in *Proc. IEEE ICASSP*, vol. 2, pp. 1259–1262, 1997.