EMD-BASED NOISE ESTIMATION AND TRACKING (ENET) WITH APPLICATION TO SPEECH ENHANCEMENT

Navin Chatlani and John J. Soraghan

Centre for Excellence in Signal and Image Processing, University of Strathclyde
Royal College Building, 204 George Street, Glasgow
phone: +(0044) 141 548 2205, email: navin.chatlani@eee.strath.ac.uk, j.soraghan@eee.strath.ac.uk
http://www.eee.strath.ac.uk

ABSTRACT

Speech enhancement from measured speech signals is fundamental in a wide range of instruments. It relies on a noise estimate which can be obtained using techniques such as the minimum statistics (MS) approach. In this paper, a novel approach for Empirical Mode Decomposition (EMD) based noise estimation and tracking (ENET) is presented with application to speech enhancement. Spectral analysis of non-stationary signals such as speech is performed effectively using EMD. The Improved Minima Controlled Recursive Averaging (IMCRA) that evolved from MS has been shown to be effective in non-stationary environments. ENET is able to use EMD in a novel way to estimate the noise spectrum more accurately than IMCRA and enhance speech more effectively than conventional log-MMSE approaches. A comparative performance study is included that demonstrates that it achieves improved speech quality than a conventional log-MMSE filtering approach with better noise estimation, even during periods of strong speech activity.

1. INTRODUCTION

A common problem encountered in speech enhancement systems is the removal of unwanted disturbances, i.e. noise from desired speech signals. Adaptive noise cancellation is commonly performed when enhancing speech sequences when a noise reference is available. Single-channel speech enhancement systems traditionally employ Voice Activity Detection (VAD) to estimate the statistics of the noise signal during silent segments. Newer flavours of noise estimation systems such as the MS-based [1] approaches and IMCRA [2] are able to estimate the noise spectrum based on the observation that the noisy signal power decays to values characteristic of the contaminating noise during speech pauses. Significant interest is given to speech enhancement systems that have developed from the log-spectral amplitude estimator (LSA) [3].

Empirical Mode Decomposition (EMD) is an effective multi-resolution approach for analyzing non-stationary signals such as speech. By performing a sifting process, the EMD decomposes the desired signal into Intrinsic Mode Functions (IMFs) which are data-adaptive as opposed to other transforms such as the Discrete Wavelet Transform (DWT) which use predefined basis functions. Recent approaches for dual-channel [4] and single-channel speech enhancement [5][6][7] using EMD have been developed. The EMD-based denoising [5] and EMD-MMSE [6] of signals contaminated with stationary white noise are based on an empirically observed noise model derived from a study of IMF statistics in noise-only situations. Denoising involved removal of those IMFs whose energies exceeded a predefined threshold and EMD-MMSE was performed by filtering the IMFs formed from the decomposition of speech contaminated with white Gaussian noise. In [7], an optimum gain function is estimated for each IMF to suppress residual noise that may be retained after single channel speech enhancement algorithms.

In this paper, a new ENET technique is proposed for speech enhancement and noise estimation. ENET uses EMD, IMCRA and the LSA estimator to provide improved noise estimation and speech enhancement. The background necessary to understand the EMD and IMCRA is first presented in sections 2 and 3 respectively. In section 4, the novel ENET system with application to speech enhancement is developed. In section 5, results obtained from testing and comparing ENET to basic IMCRA/LSA speech estimation are presented and discussed. These tests are performed in non-stationary and varying SNR conditions. It shows that ENET has significant potential when performed in highly non-stationary and time-varying environments. It also demonstrates the improved noise tracking under strong speech presence. Finally, conclusions are made in section 6.

2. EMPIRICAL MODE DECOMPOSITION

2.1 Background

EMD [8][9] is a non-linear technique for analyzing and representing non-stationary signals. EMD is data-driven and decomposes a time domain signal into a complete and finite set of adaptive basis functions which are defined as Intrinsic Mode Functions (IMFs). Although these IMFs are not predefined as is the case with the Fourier and the Wavelet Transforms, the IMFs that are extracted are oscillatory and have no DC component. Figure 1 illustrates the main stages in the EMD algorithm. EMD examines the signal between two consecutive extrema (e.g. minima) and picks out the high frequency component that exists between these two points. The remaining local, low frequency component can then be found. The motivation behind the EMD is to perform this procedure on the entire signal and then to iterate on the residual low frequency parts. This allows identification of the

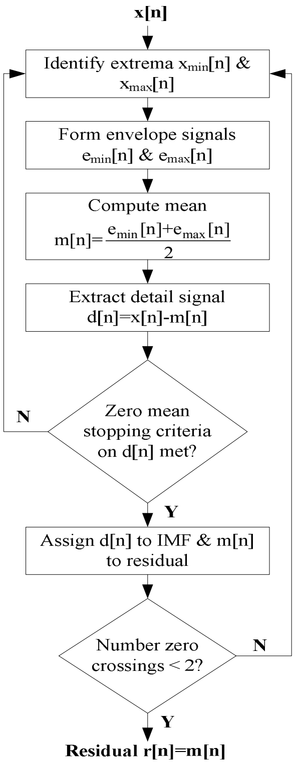

x[n]
Identify extrema $x_{min}[n]$ & $x_{max}[n]$
Form envelope signals $e_{min}[n]$ & $e_{max}[n]$
Compute mean $m[n]=\frac{e_{min}[n]+e_{max}[n]}{2}$
Extract detail signal d[n]=x[n]-m[n]
Zero mean stopping criteria on d[n] met?
N
Y
Assign d[n] to IMF & m[n] to residual
Number zero crossings < 2?
N
Y
Residual r[n]=m[n]

Figure 1 – EMD algorithm

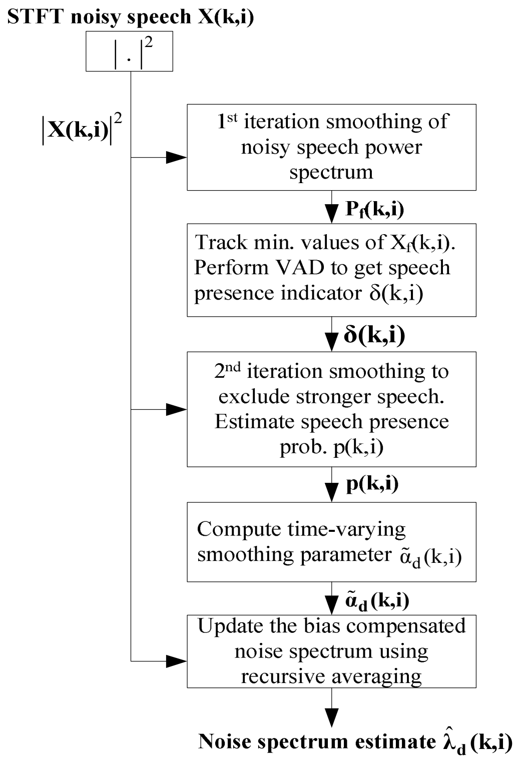

STFT noisy speech X(k,i)
$|\cdot|^2$
$|X(k,i)|^2$
1st iteration smoothing of noisy speech power spectrum
$P_f(k,i)$
Track min. values of $X_f(k,i)$. Perform VAD to get speech presence indicator δ(k,i)
δ(k,i)
2nd iteration smoothing to exclude stronger speech. Estimate speech presence prob. p(k,i)
p(k,i)
Compute time-varying smoothing parameter $\tilde{\alpha}_d(k,i)$
$\tilde{\alpha}_d(k,i)$
Update the bias compensated noise spectrum using recursive averaging
Noise spectrum estimate $\hat{\lambda}_d(k,i)$

Figure 2 – Block diagram of IMCRA noise estimation

different oscillatory modes that exist in the signal. The IMFs found must be symmetric with respect to local zero means and have the same number of zero crossings and extrema. The IMF is considered as zero-mean based on some stopping criteria such as the standard deviation between consecutively sifted functions [9].

By use of the EMD, the frequency information is embedded in the IMFs. These data-adaptive basis functions give physical meaning to the underlying process. The reconstruction process is given in (1), which involves combining the N IMFs and the residual r[n]:

$$x[n] = \sum_{i=1}^{N} IMF[n] + r[n] \qquad (1)$$

3. IMPROVED MINIMA CONTROLLED RECURSIVE AVERAGING

IMCRA combines the minimum statistics approach with recursive averaging to perform noise spectrum estimation. A summary of the IMCRA algorithm is shown in Figure 2. Consider the model described by:

$$x[n] = s[n] + d[n] \qquad (2)$$

where $x[n]$ is the noisy speech signal, $s[n]$ is the original noise-free speech, and $d[n]$ is the noise source. Assuming the independence of the speech and the noise, the STFT of (2) gives:

$$X(k,i) = S(k,i) + D(k,i) \qquad (3)$$

for frequency bin *k* and time frame *i*. It is assumed that the STFT coefficients of both the speech and the noise have complex Gaussian distributions [3]. The first iteration of smoothing of the noisy speech spectrum $P(k,i) = |X(k,i)|^2$ is performed in frequency and time to give the smoothed power spectrum $P_f(k,i)$. The minima values of $P_f(k,i)$ are tracked using the MS approach, over a specified finite window of length D, to obtain $P_{f,\min}(k,i)$. Rough VAD is performed after smoothing and minimum tracking to produce an indicator function $\delta(k,i)$ for speech presence. This speech presence decision is based on conditions [2] set on the following ratios $\gamma_{\min}(k,i)$ and $\zeta(k,i)$ as defined by:

$$\gamma_{\min}(k,i) \triangleq \frac{P(k,i)}{B_{\min} P_{f,\min}(k,i)} \qquad \zeta(k,i) \triangleq \frac{P_f(k,i)}{B_{\min} P_{f,\min}(k,i)}$$

where $B_{\min}$ is the bias of the minimum noise estimate.

$\delta(k,i)$ is used in the second smoothing stage to eliminate strong speech components from the short term spectrum $\mathrm{P}(k,i)$ before the time-domain recursive averaging. This exclusion enables improved minima tracking among the power components primarily associated with the contaminating noise source. The speech presence probability, $p(k,i)$ is then estimated and used to compute the time- varying, frequency dependent smoothing factor $\tilde{\alpha}_d(k,i)$ [2] as shown in (4) below:

$$\tilde{\alpha}_d(k,i) = \alpha_d + (1-\alpha_d)\,p(k,i) \qquad (4)$$

where smoothing parameter α_d ranges from [0, 1].

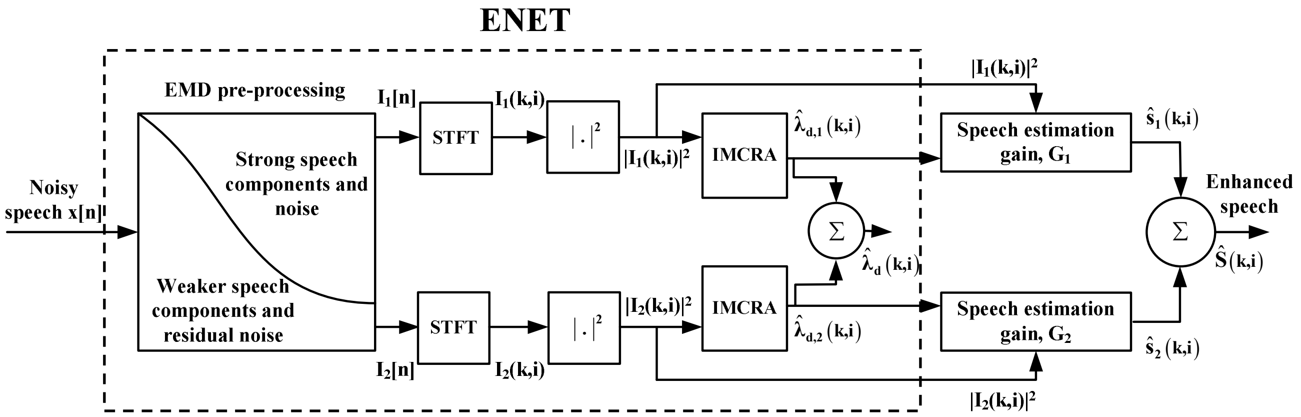


Figure 3 – Block diagram of ENET with application to speech enhancement

Recursive averaging of the power spectral values, $P(k,i)$, is then performed to obtain an estimate of the noise spectrum $\hat{\lambda}_d(k,i)$. IMCRA was shown in [2] to be more robust than the basic MS method since the minimum tracking was not used directly in the noise estimation.

4. EMD-BASED NOISE ESTIMATION AND TRACKING (ENET) WITH APPLICATION TO SPEECH ENHANCEMENT

Single channel speech enhancement algorithms rely on accurate noise spectrum estimation and speech estimation. It was shown in [2] that eliminating strong speech segments from the second smoothing stage in IMCRA improves minima tracking and the estimation of the speech presence probability. The new ENET system with application to speech enhancement is illustrated in Figure 3.

The EMD pre-processing stage from Figure 3 breaks up the signal into two bands. One band has separated the stronger speech components along with some of the noise while the other contains the weaker speech components with the residual noise. These are then processed individually to estimate the noise power components and the speech components in each signal band.

4.1 ENET Analysis

The EMD pre-processing stage in Figure 3 decomposes the signal into two signal spaces which are useful for noise tracking and speech estimation. It can be interpreted as:

$$x[n]=\sum_{j=1}^{M} IMF_j[n]+\sum_{j=M+1}^{N} IMF_j[n]=\sum_{c=1}^{2} I_c[n] \qquad (5)$$

where the EMD of the noisy signal $x[n]$ produces N IMFs and M<N. From (5), when c=1, let $I_c[n]$ denote the band that contains stronger speech components as well as some noise. When c=2, let $I_c[n]$ contain the residual noise as well as the weaker speech. Correspondingly, let $I_c(k,i)$ denote the STFT of $I_c[n]$. The IMCRA noise estimation routine is performed in ENET using the short-time power spectrum $|I_c(k,i)|^2$. The noise power spectrum, $\hat{\lambda}_{d,c}(k,i)$, in the signal band c is estimated using recursive averaging:

$$\hat{\lambda}_{d,c}(k,i+1)=\tilde{\alpha}_{d,c}(k,i)\hat{\lambda}_{d,c}(k,i)+\left[1-\tilde{\alpha}_{d,c}(k,i)\right]\left|I_c(k,i)\right|^2 \qquad (6)$$

After computing the noise estimate $\hat{\lambda}_{d,c}(k,i)$ in ENET, enhancement may be performed using the optimal LSA estimator [3]:

$$\mathrm{E}_{\min}\left[\{\lg S_c(k,i)-\lg \hat{S}_c(k,i)\}^2\right]$$

where $S_c(k,i)$ is the speech amplitude component that exists in band c and $\hat{S}_c(k,i)$ is the optimal speech estimate. The a priori SNR $\hat{\xi}_c(k,i)$ is estimated using the modified, decision directed approach in [10]. The corresponding LSA gain function is derived in [3] and denoted as $G_{LSA,c}(k,i)$ for band c:

$$G_{LSA,c} \triangleq \frac{\hat{\xi}_c(k,i)}{1+\hat{\xi}_c(k,i)}\exp\left(\frac{1}{2}\int_{v_c(k,i)}^{\infty}\frac{e^{-t}}{t}dt\right) \qquad (7)$$

where $v_c(k,i)$ is a function of the a priori and a posteriori SNR as shown in [3]. The optimally modified LSA (OM-LSA) estimator defined in [10] incorporates speech presence uncertainty to produce the gain $G_c(k,i)$:

$$G_c(k,i)=G_{LSA,c}(k,i)^{p_c(k,i)}\,G_{\min}{}^{1-p_c(k,i)} \qquad (8)$$

where $p_c(k,i)$ is the conditional speech presence probability in band c, and the threshold $G_{\min}$ is obtained from [10] based on subjective criteria. Let $\mathbf{I}=\left[I_1(k,i)\;\; I_2(k,i)\right]$ and let $\mathbf{G}=\left[G_1(k,i)\;\; G_2(k,i)\right]$. The enhanced speech signal is then estimated from the noisy signal by:

$$\hat{S}(k,i)=\mathbf{G}\mathbf{I}^T \qquad (9)$$

5. PERFORMANCE EVALUATION

The ENET technique from Figure 3 was tested on speech signals corrupted with different types of non-stationary background noises. Its performance was compared with the standard IMCRA algorithm for estimating the noise power spectrum as well as enhancement of the noisy speech. A sampling frequency of 16 kHz was used and the parameters used for IMCRA were given in [2]. The signal was split up into frames of length 512 samples and a window overlap factor of 50%. A speech utterance was obtained from the TIMIT [11] database and degraded with F16 cockpit noise and car interior noise. For these results, a speech utterance of length 40,000 samples was used to allow close examination and comparison of ENET to the basic IMCRA. These noise sources were obtained from the Noisex92 [12] database.

The superiority of ENET at noise estimation is demonstrated in Figure 4. A speech signal corrupted with car interior noise was input into both the IMCRA and the ENET algorithms at SNR level of -10 dB. Under these heavy noise conditions, both methods perform good noise tracking at frequency bin 5 (k=5) as shown by their periodograms in Figure 4(a). However, examinations of other parts of the spectrum reveal that the basic IMCRA is unable to track changes in the noise spectrum. An example of this is shown in Figure 4(b) for the first bin (k=1). It is known that this bin contains large speech spectral peaks and therefore the basic IMCRA does not update the noise spectrum. This occurs when the ratio of the speech power to the noise power is large indicating a high probability of speech presence, p(k,i) from (4) and (6). However, due to the separation of the strong speech segments, the ENET technique enables improved tracking of the noise and therefore provides a better noise estimate.

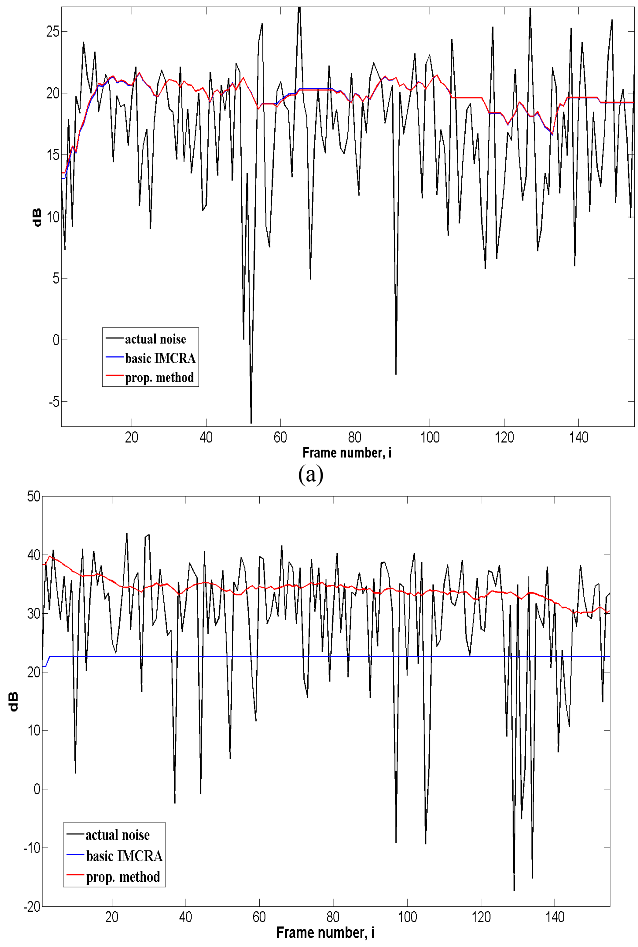
Quantitatively, noise estimation algorithms may be compared by computing the relative estimation error evaluated over N frames as given by:

$$\text{Error}_{est} = \frac{1}{N}\sum_{i=0}^{N-1}\frac{\sum_{k}\left[\hat{\lambda}_d(k,i)-\lambda_d(k,i)\right]^2}{\sum_{k}\left[\lambda_d(k,i)\right]^2} \tag{10}$$

where $\hat{\lambda}_d(k,i)$ is the noise estimated from the algorithm being tested and $\lambda_d(k,i)$ is the ideal noise spectrum. Table 1 shows the $\text{Error}_{\text{est}}$ obtained for both methods tested under various SNR levels and different non-stationary noise types. It is clear that ENET performs better estimation as it achieves a significantly lower estimation error, especially in the case of car interior noise which has a low-pass characteristic.

The OM-LSA speech estimator (8) was used to perform enhancement of the noisy speech output from both the basic IMCRA and the ENET system. In order to assess the relative performance of this new approach for speech enhancement, different subjective and objective assessment measures can be used. Subjective measures include Mean Opinion Score (MOS) and Diagnostic Acceptability Measure (DAM). Objective measures include the widely used Segmental SNR (SegSNR), Weighted Slope Spectral (WSS) distance, Log-Likelihood Ratio (LLR) and the more recent Perceptual Evaluation of Speech Quality (PESQ). However, recent studies in [13] revealed that the SegSNR does not have as high a correlation with signal and overall quality when compared to other measures such as LLR and PESQ. Therefore, the rating from the composite measure [13] for measuring overall quality (C_{OVL}) which accounts for both signal and background noise distortion will be presented for evaluation of ENET. This quality index lies in the range [1 5] and is given by:

$$C_{OVL}=1.594+0.805PESQ-0.512LLR-0.007WSS \tag{11}$$



(a)

(b)

Figure 4 – Comparison of noise estimation periodograms using basic IMCRA and ENET methods for speech contaminated with car interior noise at -10 dB. (a) Tracking at freq. bin k=5 (b) Tracking at freq. bin k=1

Table 2 presents examples of the improved quality of speech enhancement obtained by the ENET approach for speech enhancement as opposed to the previous full-band IMCRA method. These enhancement results were obtained from speech signals contaminated with F16 and car interior noise respectively. They demonstrate that the full-band approach for enhancement is inferior to ENET in low SNR, non-stationary adverse conditions. Listeners appear to be particularly sensitive to speech distortion [13] and it was found that the new algorithm gives the desired significant improvement. This occurs due to the separation of the noisy speech which allows better speech estimation in each of the two bands that are formed.

Estimation Error

Input SNR (dB)	F16		Car Interior Noise	
	IMCRA	ENET	IMCRA	ENET
8	4.93	4.54	90.67	39.37
4	4.55	4.05	91.16	1.28
0	4.38	3.79	91.57	5.63
-4	4.28	3.65	91.83	0.97
-8	4.22	3.60	91.99	0.77

Table 1 – Relative estimation error ($Error_{est}$) for noise tracking for basic IMCRA and the ENET methods under varying SNR conditions and noise sources

Composite Overall rating

Input SNR (dB)	F16		Car Interior Noise	
	IMCRA	ENET	IMCRA	ENET
10	2.96	3.01	4.20	4.35
6	2.56	2.67	3.96	4.01
2	2.19	2.28	3.67	3.82
0	1.97	2.05	3.50	3.65
-2	1.70	1.82	3.33	3.52
-6	1.20	1.29	3.01	3.29
-10	1.00	1.00	2.65	2.88

Table 2 – Comparison of signal quality using the composite overall rating (C_{OVL}) under varying SNR conditions and noise sources

The ENET approach was shown to be more effective than conventional techniques since it yields improved noise estimation and superior quality of the enhanced speech. The frequency characteristics of the contaminating noise determine the value of M from (5). This allows the separation of the signal space into the two bands of stronger speech components with noise and weaker speech components with residual noise. In the case of the lower frequency car interior noise, M was selected as 8. The F16 noise had dominant spectral peaks at higher frequencies and therefore a value of 2 was chosen for M. Although $Error_{est}$ was used for the quantitative assessment of the noise tracking, there are some deficiencies in using MSE for evaluating relative estimation performance. MSE is sensitive to outliers and also does not place relevant emphasis on over-estimation and under-estimation errors, which have different consequences for the speech estimation [1].

6. CONCLUSION

The basic IMCRA technique is effective at updating the noise spectrum by applying recursive averaging. However, the results demonstrate that the new ENET system performs better at noise tracking and also provides lower estimation errors. This technique has also been shown to yield improvements for speech enhancement systems by providing superior signal quality.

The time-varying, frequency-dependent smoothing factor used in recursive averaging during noise estimation varies with speech presence probability, and thus allows noise spectral updates even during speech activity. However, when the ratio of the speech power to the noise power is large, the estimation procedure is unable to track changes in the noise spectrum. As seen from the analysis presented, after decomposing the signal into its IMFs using the EMD, ENET is able to separate the signal space into two bands. One band has stronger speech components and noise and the other has weaker speech components and residual noise. Integration of this new approach into noise estimation and speech enhancement systems has shown that it can provide better performance for noise tracking and speech enhancement, in non-stationary and low SNR conditions.

REFERENCES

[1] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," IEEE Trans. on Speech and Audio Processing, vol. 9, no. 5, pp. 504-512, Jul 2001.

[2] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," IEEE Trans. on Speech and Audio Processing, vol. 11, no. 5, pp. 466-475, Sept 2003.

[3] Y. Ephraim, D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," IEEE Trans. on Acoustics, Speech and Signal Processing, vol. 33, no. 2, pp. 443-445, Apr 1985.

[4] N. Chatlani, J. J. Soraghan, "Adaptive Empirical Mode Decomposition for Signal Enhancement with application to speech," 15th Int'l Conference on Systems, Signals and Image Processing 2008, pp. 101-104, June 2008.

[5] P. Flandrin, P. Goncalves and G. Rilling, "Detrending and Denoising with Empirical Mode Decompositions", in Proc. EUSIPCO 2004, pp. 1581-1584, 2004.

[6] K. Khaldi, A. O. Boudraa, A. Bouchikhi, and M. T-H Alouane, "Speech Enhancement via EMD", in EURASIP Journal on Advances in Signal Processing, vol. 2008, Article ID 873204, 8 pages, 2008.

[7] T. Hasan, M. K. Hasan, "Suppression of Residual Noise From Speech Signals Using Empirical Mode Decomposition," Signal Processing Letters, IEEE , vol. 16, no. 1, pp. 2-5, Jan 2009.

[8] N. E. Huang et al., "The Empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," in Proc. Royal Society London A, vol. 454, pp. 903-995, 1998.

[9] G. Rilling, P. Flandrin and P. Goncalves, "On empirical mode decomposition and its algorithms", in IEEE-EURASIP Workshop NSIP, Jun 8-11, 2003.

[10] I. Cohen and B. Berdugo, "Speech enhancement for nonstationary noise environments," Signal Process., vol. 81, no. 11, pp. 2403-2418, Nov 2001.

[11] TIMIT speech database, Speech Enhancement and Assessment Resource,
<http://cslu.cse.ogi.edu/nsel/data/SpEAR_noisyspeech.html> [accessed Feb 2009]

[12] NOISEX – 92 Database,
<http://spib.rice.edu/spib/select_noise.html> [accessed Feb 2009]

[13] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement", in IEEE Trans. on Audio, Speech and Lang. Processing, vol. 16, no. 1, Jan 2008.