EMPIRICAL SIGNAL DECOMPOSITION FOR ACOUSTIC NOISE DETECTION

*L. Zão and R. Coelho**

Laboratory of Acoustic Signal Processing (LASP)
Military Institute of Engineering (IME)
Rio de Janeiro, Brazil
{zao, coelho}@ime.eb.br

ABSTRACT

This paper introduces an adaptive noise detection method for non-stationary acoustic noisy signals. The proposed approach is based on the empirical mode decomposition (EMD) and a vector of Hurst exponent coefficients. The scheme is investigated considering real acoustic noisy signals with different non-stationarity degree and signal-to-noise ratio (SNR). The results demonstrate that the EMD-based noise detector enables a better separation between the clean and noisy signals when compared to the competing methods. It also leads to an average SNR improvement of 4.4 dB for the resulting enhanced signals.

***Index Terms*—** empirical mode decomposition, Hurst exponent, acoustic noise detection, index of non-stationarity.

1. INTRODUCTION

Wireless technology is a reality in nowadays professional and personal lives. In the last years, smartphones, tablets and laptops powered with advanced acoustic sensors, processors, and high-speed connection became very popular. This emerging scenario boosted the signal processing research to support the fast growing needs of the diversity of applications that underline these devices. Moreover, a broad set of these applications is applied in real acoustic noisy environment (restaurant, street, traffic, train). Consequently, noise detection is a main issue to proceed the acoustic scene analysis, signal processing and classification. Furthermore, real acoustic signals are generally nonlinear and non-stationary.

The empirical mode decomposition (EMD) [1] is a new concept and a powerful tool for signal processing in time domain. The data-driven method is devoted to treat nonlinear and non-stationary sequences. Its multiresolution analysis decomposes a signal into a series of oscillatory intrinsic mode functions (IMF) and a residual component. Different from the traditional wavelets, a set of *a priori* fixed basis functions is not required for the decomposition process. Instead, the IMFs are completely based on the local properties of the input data, which guarantees its adaptivity to any kind of signal. In [2], the authors showed that when applied to fractional Gaussian noises, EMD behaves like a dyadic filterbank with overlapping band-pass filters. Due to these features, EMD is being investigated in many areas, e.g., biomedical, meteorological, ocean engineering, and seismic.

This work proposes a scheme to detect noise components from a corrupted acoustic signal collected in a real environment with different non-stationarity degree. In this proposal, the EMD is firstly applied to the noisy signal. Then, the noise components of each IMF are identified on a frame-by-frame basis by a vector of Hurst scaling exponent [3]. The extraction of those detected noise components enables further reconstruction of a target signal.

The evaluation experiments are conducted considering real target signals and acoustic noises with different indices of non-stationarity (INS) [4]. Two baseline detection methods are considered for the investigation of the proposed approach: variance [5] and standardized mean [6]. The Hurst-based noise detection also leads to substantial quality improvement for most of the noisy scenarios. The results show that the proposed method outperforms the baseline in terms of signal-to-noise ratio and segmental SNR (SegSNR).

2. ACOUSTIC NOISE DETECTION METHOD

The first step of the proposed method refers to the decomposition of the noisy signal. Based on the Hilbert-Huang transform (HHT) [1], EMD locally analyzes a signal $x(t)$ between two consecutive extrema (minima or maxima). While a fast oscillation (high-frequency) defines a detail function, the remaining slow oscillation (low-frequency) indicates a local trend or residual. By definition, an IMF has zero mean and all its local maxima and minima are positive and negative, respectively. The first detail function, $d_1(t)$, is obtained from all the consecutive extrema of $x(t)$, such that $x(t) = d_1(t) + a_1(t)$, where $a_1(t)$ denotes the first local trend. In general, the separation between the fast and slow oscillations is repeated over the residual of order $k-1$ to obtain the detail

*This work was partially supported by the National Council for Scientific and Technological Development (CNPq) under 307866/2015-7 research grant.

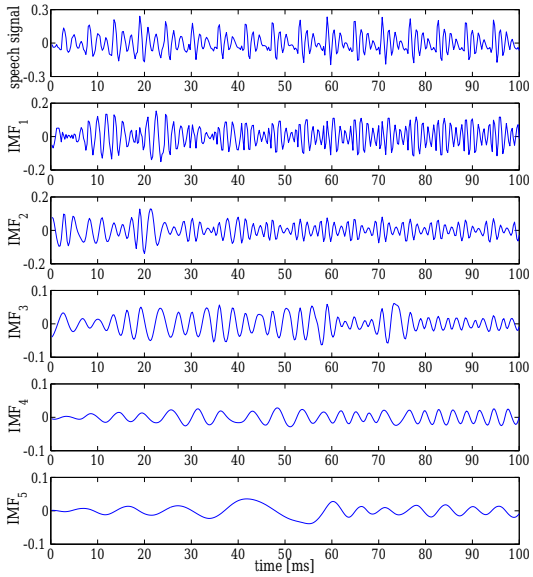


Fig. 1. Amplitude of a clean speech signal and the corresponding five IMFs.

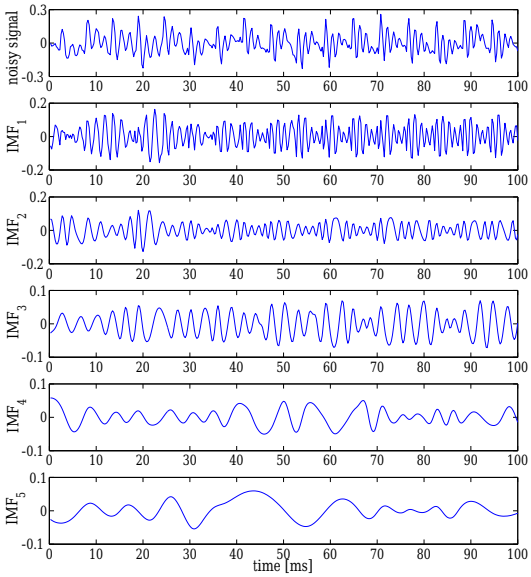


Fig. 2. Amplitude of a noisy speech signal corrupted with car traffic noise and the corresponding five IMFs.

and the local trend of order k, i.e., $a_{k-1}(t) = d_k(t) + a_k(t)$. The decomposition stops when the current residual may no longer be decomposed into new IMFs.

2.1. EMD algorithm

The EMD algorithm can be described as follows:

1. Set $k = 1$ and initialize the variable $a_0(t) = x(t)$;
2. Identify all local minima and maxima of $a_{k-1}(t)$;
3. Obtain the upper ($e_{max}(t)$) and lower ($e_{min}(t)$) envelopes by cubic splines interpolation of the local maxima and minima, respectively;
4. Compute the local trend as the average between the envelopes, i.e., $a_k(t) = (e_{min}(t) + e_{max}(t))/2$;
5. Calculate $d_k(t) = a_{k-1}(t) - a_k(t)$ as the new detail function;
6. Set $k = k + 1$ and iterate steps 2-5 on the new residual local trend $a_k(t)$.

If a detail function $d_k(t)$, obtained in step 5, does not follow the IMF definition, steps 2 to 5 are repeated with $d_k(t)$ in place of $a_{k-1}(t)$. This process, called *sifting*, is repeated until a new $d_k(t)$ can finally be considered as an IMF. The EMD algorithm assures completeness of the analyzed signal, i.e., $x(t) = \sum_{k=1}^{K} \text{IMF}_k(t) + r(t)$, where K is the total number of IMFs, $\text{IMF}_k(t)$ denotes the k-th mode and $r(t) = a_K(t)$ is the last residual. A detailed description of the EMD method, its refined version and application to signal enhancement can be found in [7].

Figs. 1 and 2 illustrate an example of the first five IMFs obtained from the decomposition of a clean and a noisy speech signal, respectively. The clean signal corresponds to a speech segment collected from the TIMIT [8] database. A real car traffic[1] noise is selected to corrupt the speech segment with SNR of 5 dB. Note that the first mode is composed of faster oscillations when compared to the second IMF, which has faster oscillations than the third one, and so on. It can also be seen from Figs. 1 and 2 that the first two IMFs are quite similar in the clean and the noisy signal. However, the effects of the acoustic noises are noticeable after the third IMF. This is an indication that EMD is suitable for the noise components detection.

2.2. Noise Components Identification

The second step of the detection scheme consists in the identification of the noise components of each IMF by the estimation of a vector of Hurst exponent coefficients. The main goal is to define an IMF index L ($1 \leq L \leq K$) such that the noise components are concentrated at IMFs with indices $k \geq L$. The Hurst exponent ($0 \leq H \leq 1$) was chosen for the IMFs selection since it expresses the scaling degree of a signal and is related to its power spectral characteristics, i.e, it can be denoted as a time-frequency parameter. For example, if $x(t)$ is a white noise, its power spectral density is approximately constant and $H = 1/2$. When low frequencies are prominent, then $H > 1/2$. If the energy is mostly concentrated at the high frequencies then $H < 1/2$. Due to such characteristics, the Hurst exponent was proposed in [9] to compose a speech feature vector and successfully applied to speaker recognition. In this work, the wavelet-based estimator [10] was adopted to obtain the H values of the IMFs on a frame-by-frame basis.

[1] Available at http://www.freesound.org.

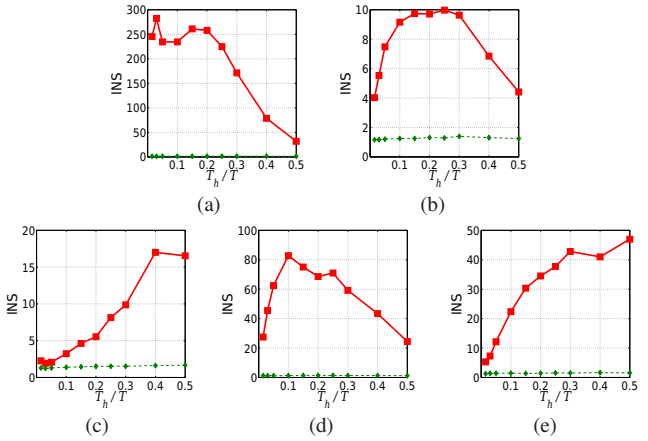


Fig. 3. Index of non-stationarity of acoustic signals: (a) speech, (b) babble, and noises: (c) car traffic, (d) chainsaw, and (e) jackhammer. The dashed lines indicate the values of the threshold γ for the stationarity tests.

Since the proposed detection scheme is mainly concerned with non-stationary signals, the index of non-stationarity (INS) [4] is adopted to objectively examine and quantify the non-stationarity degree of the acoustic noises and target signals. The stationarity test is conducted by comparing the spectral components of the signal to a set of stationary references called *surrogates*. For each window length T_h, a threshold γ is defined for the stationarity test. Thus, the signal is considered as non-stationary if INS $> \gamma$.

In this work, three acoustic noises (car traffic, chainsaw[2] and jackhammer[2]) are used to corrupt two target signals: speech (the same used in Figs. 1 and 2) and babble[3]. The signals and noises have 3 s time duration and are sampled at 8 kHz. The INS values obtained from these signals and noises are depicted in Fig. 3 (continuous lines). The time scale is the ratio of the length of the short-time spectral analysis (T_h), and the total time duration ($T = 3$ s) of the sample sequences. From the INS results, the target signals and noises are here defined as: speech signal and chainsaw noise are highly non-stationary (HNS; INS ≥ 80); jackhammer noise is non-stationary (NS; INS ≥ 40); babble signal and car traffic noise are moderately non-stationary (MNS; INS < 20). The INS of the speech signal corrupted with the three noises with SNR of 5 dB are shown in Fig. 4. It can be seen that these noisy signals are highly non-stationary, i.e., INS ≥ 80, even for the car traffic noise (MNS).

Fig. 5 shows the H average values estimated from IMFs segments of 32 ms (rectangular window with 256 samples). The H results of the target clean speech signal are presented in the dashed lines. The continuous lines correspond to the speech signals corrupted by the car traffic and jackhammer noises with SNR of 5 dB. It can be observed that the noise

[2]Available at http://www.freesfx.co.uk.

[3]Collected from the NOISEX database [11].

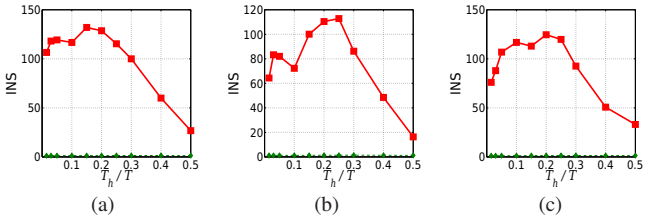


Fig. 4. Index of non-stationarity of the speech signal corrupted with the acoustic noises with SNR of 5 dB: (a) car traffic, (b) chainsaw, and (c) jackhammer.

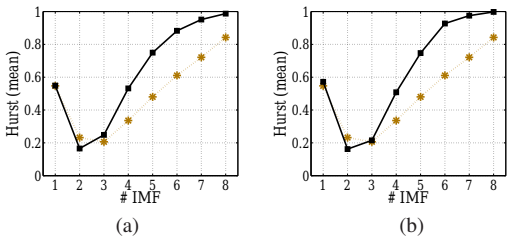


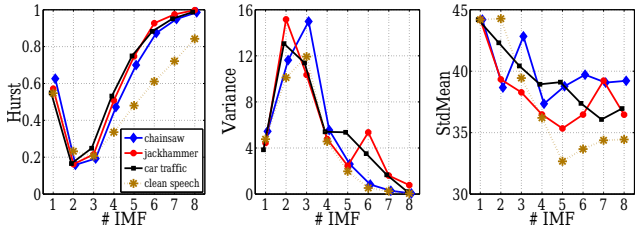
Fig. 5. H values estimated from the IMFs corresponding to speech signal corrupted by acoustic noises (continuous lines) with SNR of 5 dB: (a) car traffic and (b) jackhammer. Dashed lines refer to the clean speech.

components differ from the clean speech after the fourth IMF. As expected, this corresponds to $H > 1/2$, i.e., low-frequency noise. This demonstrates that the Hurst exponent is a good criterion for the acoustic noise detection.

Additionally, for the evaluation of the proposed noise detection scheme, the variance [5] and standardized mean (StdMean) [6] are adopted as baseline criteria. The variance-based criterion identifies the first IMF with index L ($L \geq 4$) where the variance is greater than the adjacent modes, i.e., $\mathrm{Var}\left[\mathrm{IMF}_L(t)\right] > \mathrm{Var}\left[\mathrm{IMF}_{L-1}(t)\right]$ and $\mathrm{Var}\left[\mathrm{IMF}_L(t)\right] > \mathrm{Var}\left[\mathrm{IMF}_{L+1}(t)\right]$. The noise components are assumed to be concentrated at the IMFs with index $k \geq L$. The StdMean is estimated from each IMF and is defined as the ration of the mean and the standard deviation. In this work, the StdMean-based criterion identifies L as the first index for which StdMean is greater than the root mean square of the standardized mean of the first four modes.

3. SIGNAL RECONSTRUCTION

Since the noise components are assumed to be mostly concentrated at the IMFs with $H > 1/2$ (low-frequency), the target signal reconstruction will be composed with the remaining modes ($H < 1/2$). After the decomposition of the noisy signal, each IMF is divided into Q short-time frames. For each frame q, the Hurst exponent defines an index L such that the target signal frame $\hat{x}_q(t)$ is reconstructed using only the first $L-1$ modes., i.e., $\hat{x}_q(t) = \sum_{k=1}^{L-1} \mathrm{IMF}_{q,k}(t)$, where $\mathrm{IMF}_{q,k}(t)$ is the q-th frame of the k-th IMF. The target signal is finally given by the concatenation of all Q frames. In



(a) (b) (c)

Fig. 6. The average values of (a) H, (b) variance and (c) StdMean estimated from the IMFs of the clean and noisy speech signals (SNR = 5 dB).

this work, the target signals are also reconstructed with the baseline StdMean [5] and variance [6] criteria.

In order to investigate the noise interference reduction of the proposed and baseline approaches, the target signals are reconstructed and further evaluated in terms of two signal quality measures: SNR and SegSNR. SegSNR is defined as $\text{SegSNR} = \frac{10}{Q}\sum_{q=1}^{Q}\log\frac{\sum_{t=1}^{T_d}x_q^2(t)}{\sum_{t=1}^{T_d}[x_q(t)-\hat{x}_q(t)]^2}$, where Q is the total number of frames, T_d is the frame length (in samples), and $x_q(t)$ and $\hat{x}_q(t)$ represent the q-th frame of the clean and reconstructed target signals, respectively,

4. EXPERIMENTS AND RESULTS

In this Section, the Hurst, variance and StdMean approaches are compared for acoustic noise perception. The IMFs selection and the signal reconstruction are performed with frame length of 32 ms. Fig. 6 illustrates the noise detection obtained with the Hurst exponent vector and the baseline criteria. The H, variance and StdMean average values are estimated from each IMF of the clean and noisy versions of the speech signal, considering SNR of 5 dB. Note from Fig. 6(a) that the H values of the clean and noisy signals are clearly different after the third IMF. This indicates that the proposed method is able to detect the noise components from the three noisy signals. From Fig. 6(b) it can also be seen that variance peaks appear at the sixth and fifth modes for the jackhammer (NS) and car traffic (MNS) noises, respectively. However, there is no variance peak after the fourth IMF for the chainsaw noise (HNS), which can lead to a noise detection error. Finally, the StdMean values in Fig. 6(c) indicate that the noise components can only be identified after the fifth IMF for any noisy signal. These results reinforce that the separation of the clean speech and noisy signals is clearly more evident with the proposed detection method, even for the chainsaw noise (HNS).

The SNR and SegSNR are here examined for the target signal reconstruction with the proposed and baseline noise detection methods. The target signals are corrupted considering three SNR values: -5 dB, 0 dB and 5 dB. Tables 1 and 2 present the SNR and the SegSNR improvement results computed from the reconstructed signals using the Hurst, variance and StdMean noise detection criteria. The SegSNR is computed with frame length of 25 ms. The highlighted values correspond to signal quality improvement greater than 1 dB. The speech and babble target signals are considered in these experiments. The SNR results of the proposed scheme substantially outperforms the baseline methods for most of the noise conditions. The best SNR values are obtained for the car traffic noise (MNS), where the SNR improvement achieves 11.7 dB for the target speech signal corrupted with SNR of -5 dB. For this same condition, the StdMean and variance criteria achieve SNR gain of 0.8 dB and 1.0 dB, respectively. The Hurst-based solution provides the best SNR results even for the highly non-stationary chainsaw noise. Considering the SegSNR measure, the Hurst-based solution reaches the best results for all the noisy conditions. For example, a SegSNR gain of 5.2 dB is achieved for the speech signal corrupted with the car traffic with SNR of -5 dB. These results emphasize that the proposed scheme is suitable for the detection of real non-stationary acoustic noises.

Table 1. SNR (dB) obtained with the noise selection criteria.

		speech signal			babble signal		
Noise	SNR	Hurst	Variance	StdMean	Hurst	Variance	StdMean
car traffic (MNS)	-5	**6.7**	-4.2	-4.0	**2.5**	-4.9	-4.0
	0	**9.9**	2.2	1.4	**3.7**	0.6	1.1
	5	**12.9**	7.5	6.7	5.3	4.4	4.6
jackhammer (NS)	-5	**3.8**	-0.4	-1.5	**1.6**	-2.3	-3.1
	0	**7.6**	5.6	3.2	**4.3**	2.4	1.5
	5	**11.7**	9.7	7.8	**6.7**	5.1	5.4
chainsaw (HNS)	-5	**-3.4**	-4.7	-4.7	-4.3	-4.9	-4.8
	0	0.8	0.4	0.3	0.0	-0.4	-0.3
	5	5.3	5.2	5.2	**4.9**	3.7	3.8

Table 2. SegSNR improvement (dB) obtained with the noise selection criteria.

		speech signal			babble signal		
Noise	SNR	Hurst	Variance	StdMean	Hurst	Variance	StdMean
car traffic (MNS)	-5	**5.2**	1.2	0.9	**4.8**	0.3	0.9
	0	**4.1**	1.8	1.1	**2.2**	0.5	0.7
	5	**2.4**	1.4	1.0	0.0	-0.5	-0.4
jackhammer (NS)	-5	**4.6**	3.6	2.4	**4.7**	3.1	2.3
	0	3.4	3.3	1.8	3.0	2.1	1.4
	5	2.2	1.9	1.1	0.9	0.1	0.2
chainsaw (HNS)	-5	0.7	0.3	0.2	0.2	0.0	0.0
	0	0.3	0.2	0.2	0.0	-0.6	-0.3
	5	0.1	0.1	0.1	0.2	-1.1	-0.8

5. CONCLUSION

This paper presented a time-domain noise detection scheme for signals corrupted by non-stationary acoustic noise. The proposal is derived from a two steps procedure composed by the empirical mode decomposition and a Hurst exponent vector. The results demonstrated that the proposed method is suitable for non-stationary noise detection. A SNR gain of 1.6 dB is obtained for the highly non-stationary chainsaw noise source. Moreover, it can be very promising for speech enhancement solutions [5, 7, 12–14].

6. REFERENCES

[1] N. Huang, Z. Shen, S. Long, M. Wu, H. Shih, Q. Zheng, N. Yen, C. Tung, and 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. Series A: Mathematical, Physical and Engineering Sciences*, vol. 454, pp. 903–995, March 1998.

[2] P. Flandrin, G. Rilling, and P. Gonçalvès, "Empirical mode decomposition as a filter bank," *IEEE Signal Processing Letters*, vol. 11, pp. 112–114, February 2004.

[3] E. Hurst, "Long-term storage capacity of reservoirs," *Transactions of the American Society of Civil Engineers*, pp. 770–799, April 1951.

[4] P. Borgnat, P. Flandrin, P. Honeine, C. Richard, and J. Xiao, "Testing stationarity with surrogates: A time-frequency approach," *IEEE Transactions on Signal Processing*, vol. 58, pp. 3459–3470, July 2010.

[5] 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, pp. 1158–1166, May 2012.

[6] P. Flandrin, P. Gonçalves, and G. Rilling, "Detrending and denoising with empirical mode decompositions," *Proceedings of the European Signal Processing Conference (EUSIPCO 2004)*, pp. 1581–1584, September 2004.

[7] R. Coelho and L. Zão, "Empirical mode decomposition theory applied to speech enhancement," in *Signals and Images: Advances and Results in Speech, Estimation, Compression, Recognition, Filtering and Processing* (R. Coelho, V. Nascimento, R. Queiroz, J. Romano, and C. Cavalcante, eds.), Boca Raton, Florida: CRC Press, 2015.

[8] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium, Philadelphia*, 1993.

[9] R. Sant Ana, R. Coelho, and A. Alcaim, "Text-independent speaker recognition based on the hurst parameter and the multidimensional fractional brownian motion model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 931–940, May 2006.

[10] D. Veitch and P. Abry, "A wavelet-based joint estimator of the parameters of long-range dependence," *IEEE Transactions on Information Theory*, vol. 45, pp. 878 –897, April 1999.

[11] A. Varga and H. 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 Communications*, vol. 12, pp. 247–251, July 1993.

[12] K. Khaldi, A. Boudraa, A. Bouchikhi, and M. Alouane, "Speech enhancement via EMD," *EURASIP Journal on Advances in Signal Processing*, vol. 2008, pp. 873204.1–873204.8, May 2008.

[13] T. Hasan and M. Hasan, "Suppression of residual noise from speech signals using empirical mode decomposition," *IEEE Signal Processing Letters*, vol. 16, pp. 2–5, January 2009.

[14] L. Zão, 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, pp. 899–911, May 2014.