Speech Enhancement with Nonstationary Acoustic Noise Detection in Time Domain

R. Tavares and R. Coelho, *Member, IEEE*

***Abstract*—This letter proposes a new time domain speech enhancement technique for signals corrupted by nonstationary acoustic noises. In this method, the noise components are detected and attenuated directly from the corrupted speech samples. They are obtained with a robust estimation of the noise standard deviation considering any speech and noise amplitude distribution. These values are used to define a noise selection threshold. Additionally, this solution does not require the usage of any spectral analysis or temporal decomposition as a pre-processing phase. The experiments results show that the proposed scheme leads to significant improvement in the speech quality and intelligibility when compared to competing enhancement approaches.**

***Index Terms*—Index of nonstationarity, robust estimation, speech enhancement.**

I. Introduction

SPEECH enhancement has been the object of many studies in the signal processing area. It also underlies a diversity of applications such as speech and speaker recognition, source localization, and acoustic emotion identification. The attenuation of the noise interference is still a major challenge for the quality and intelligibility improvement of the noisy speech signals. The main issue concerns the estimation of the noise statistics, particularly in nonstationary real environments.

Generally, speech enhancement schemes can be classified by its noise statistics estimation approach, i.e., considering the spectral or time domain. Conventional spectral solutions as the spectral subtraction [1] and the minimum mean square error [2], usually apply the short-time Fourier transform (STFT) and a voice activity detector (VAD) to estimate the noise power spectrum in regions where speech is considered absent. These algorithms can obtain satisfactory results when the acoustic noise is stationary. However, in real environments acoustic noises are nonstationary [3] and, in these cases, VAD-based noise estimators are not able to attain accurate power spectrum statistics. Alternative estimators [3]–[5] were introduced to deal with nonstationary noise. In these proposals, the noise power spectrum is updated every time frame, even during speech activity. Nevertheless, they are often inaccurate in the presence of highly nonstationary noises [6].

Methods that include the estimation of noise statistics in time domain, e.g., wavelet decomposition, were also used to enhance noisy speech signals. Their main advantage is that they avoid the use of the STFT. In [7], a post-processing filtering based on the empirical mode decomposition (EMD) [8] theory was employed to remove the residual low-frequency noise after the usage of a spectral pre-enhancement scheme. This EMD-based filtering (EMDF) improved the quality of speech signals corrupted by stationary noise. The EMDH [9], [10] technique proposed the application of EMD directly to the noisy speech samples. The most corrupted components were selected and attenuated according to the Hurst exponent (H) [11] estimated from short-time frames. EMDH showed promising speech quality and intelligibility gain for signals collected in highly nonstationary noisy environments.

This letter introduces a new time domain method to enhance speech signals corrupted by nonstationary acoustic noises. Different from the other techniques presented in the literature, this proposal does not require any time-frequency analysis procedure such as Fourier transform or temporal decomposition. Here, the acoustic noise standard deviation is estimated in time domain using an adaptation of a robust estimator [12] on a frame-by-frame basis. These values are used to define a noise selection threshold. The estimation algorithm does not need any previous knowledge of the noise amplitude distribution. Thus, the proposed approach can be applied to any kind of acoustic noise.

Extensive experiments are conducted to evaluate the speech enhancement scheme. Four objective measures are used to compare the proposed and baseline techniques: unbiased minimum mean-square error (UMMSE) [4], EMDF and EMDH. Four nonstationary acoustic noises, with different indexes of nonstationarity (INS) [13], are employed to corrupt the speech signals with signal-to-noise ratios (SNR) between -10 dB and 10 dB. The results show that the proposed method outperforms the baseline solutions in terms of speech quality and intelligibility measures.

II. Proposed Speech Enhancement Method

Speech enhancement techniques are commonly implemented in four main phases:

1) Pre-processing of the noisy signal using a time-frequency procedure, e.g., STFT/VAD, EMD or wavelet;
2) Detection or estimation of the noise statistics;
3) Selection and attenuation of the noisy components;
4) Speech signal reconstruction.

The proposed speech enhancement method is performed without the usage of any pre-processing algorithm (refer to phase 1). The noise components are detected by considering

Manuscript received July 02, 2015; revised October 15, 2015; accepted October 21, 2015. Date of publication October 26, 2015; date of current version November 02, 2015. This work was supported in part by the National Council for Scientific and Technological Development (CNPq) under Grant 304254/2012-6. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Peter K. Willett.

The authors are with the Laboratory of Acoustic Signal Processing, Military Institute of Engineering (IME), Rio de Janeiro, RJ, Brazil (e-mail: coelho@ime.eb.br).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2015.2495102

their standard deviation estimated directly from the corrupted speech samples. The estimated values are used to define the selection threshold that will compose the enhanced signal. This is attained on a frame-by-frame basis and in time domain.

The speech enhancement scheme begins with the segmentation of the noisy speech signal into a set of short-time frames, $y_q(k) = y(k + qK), k = 1, \ldots, K$, where $q \in \{0, \ldots, Q-1\}$ is the frame index, K[1] is the frame length in samples, and $y(k)$ is the noisy speech sample sequence. For each corrupted speech frame $y_q(k)$, the d-Dimensional Trimmed Estimator (DATE) [12] is adopted to obtain the noise standard deviation. Originally, DATE was defined for additive white Gaussian noise. In this work, this estimator is adapted and examined to detect the acoustic noise standard deviation considering unknown speech and noise amplitude distribution. The estimated noise standard deviation is then subtracted from each corrupted sample to compose the enhanced speech signal.

The proposed algorithm can be described in three main phases: noise standard deviation estimation, selection of the noisy components, and speech signal reconstruction.

A. Noise Detection

The noise standard deviation is estimated in two main steps:
Step 1: preparation of the noisy sample sequence.

- Initialize the detection threshold

$$\xi(\rho) = \frac{1}{2}\rho + \frac{1}{\rho}\log\left(1 + \sqrt{1 - \exp(-\rho^2)}\right), \qquad (1)$$

  where $\rho = 4$ and $\xi(\rho) = 3.4742$ for a Gaussian noise [12].
- Rearrange the noisy sequence $\{y_q(k)\}$ by the order of amplitude values as $Y_1 \leq Y_2 \leq \ldots \leq Y_K$.

Step 2: estimation of the noise standard deviation.

- Compute k_{min}: this indicates the number of samples $Y_1, Y_2, \ldots Y_{k_{min}}$ that has only noise components. The algorithm assumes that speech amplitude values are above some known lower bound and that its probability of occurrence is less than 0.5. According to the Bienayme-Chebyshev-Markov inequality, this value can be obtained by $k_{min} = K/2 - hK$, where $h = \frac{1}{\sqrt{4K(1-Q)}}$ and Q is the confidence degree, which is assumed to be equal to 95% for a Gaussian noise.
- Verify if there exists an integer $k \in \{k_{min}, \ldots, K\}$ such that: $\|Y_{k-1}\| \leq \frac{\xi(\rho)}{\lambda}\frac{\sum_{i=1}^{k}\|Y_i\|}{k} < \|Y_{k+1}\|$, where $\|\cdot\|$ is the Euclidean norm and $\lambda = \sqrt{2}\Gamma(1)/\Gamma(0.5) = 0.7979$ is an adjustment factor of the detection threshold. If so, then define $b_q = k$; otherwise, $b_q = k_{min}$.
- Calculate the standard deviation:

$$\sigma_q = \frac{[\sum_{i=1}^{b_q}\|y\|]\xi(\rho)}{\lambda b_q}.$$

Fig. 1 shows the noise standard deviation values of different short-time frames of a male speech signal corrupted by the chainsaw[2] noise with SNR of 10 dB. The σ_q results obtained with DATE and the median absolute deviation (MAD) algorithms are indicated in the blue and green lines, respectively.

[1]In this work, K is set to 512 samples, which corresponds to noisy speech frames with 32 ms duration for a sampling rate of 16 kHz.

[2]Acoustic noise collected from the Freesound.org database available at: www.freesound.org.

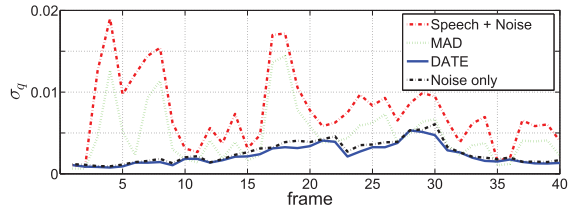

Fig. 1. Comparison between the standard deviation values estimated from a speech signal corrupted by the chainsaw noise with SNR of 10 dB ($\rho = 4$).

Note that the results estimated with DATE are much closer to the real standard deviation values of the acoustic noise. On the other hand, MAD results are similar to the standard deviation of the noisy signal. This indicates that the noise standard deviation is an interesting criteria for the noise components selection and speech signal reconstruction. These results also demonstrate that $\rho = 4$ is interesting even when applied to real acoustic noises.

B. Selection of Noisy Components

The acoustic noise standard deviation values are used to detect segments of the noisy signal where speech is considered absent. For this purpose, given the value of b_q (refer to step 2, Section II-A), the amplitude value $y(b_q)$ is defined as the threshold level to select the noise components from the corrupted signal. The amplitude values below this threshold are treated as noise only. Noise standard deviation values are then subtracted from the remaining samples to obtain the amplitudes of the enhanced speech signal.

C. Speech Signal Reconstruction

For the speech signal reconstruction, the q-th frame is composed of amplitude values given by

$$\tilde{y}_q(k) = \begin{cases} y_q(k) - \sigma_q, & \text{if } y_q(k) \geq y(b_q); \\ 0, & \text{otherwise.} \end{cases} \qquad (2)$$

where $k = 1, \ldots, K$. The enhanced speech signal $\tilde{y}(k)$ is finally achieved by concatenating all the frames obtained in (2), i.e., $\tilde{y}(k) = \sum_{q=0}^{Q-1}\tilde{y}_q(k - qK)$.

III. Experiments and Results

Extensive speech enhancement experiments are conducted with a subset of 24 speakers (16 male and 8 female) of the TIMIT speech database [14], i.e., 240 speech segments with sampling rate of 16 kHz and average time duration of 3 seconds. Four nonstationary acoustic noises are used to corrupt the speech utterances. The babble, factory chainsaw and jackhammer noises are selected, respectively, from NOISEX-92 [15] and Freesound.org[2] databases. The speech signals are corrupted considering five SNR values: -10 dB, -5 dB, 0 dB, 5 dB, 10 dB.

Fig. 2 presents the index of nonstationarity results obtained from segments of the four noises. The INS is here adopted to objectively examine the nonstationarity of the acoustic noise. The time scale T_h/T is the ratio of the length of the short-time spectral analysis (T_h) and the total time duration ($T = 3$ seconds) of the noises sample sequences. For each window length T_h, a threshold γ is defined to guarantee the stationarity assumption with a confidence degree of 95%. Thus,

$$\text{INS}\begin{cases} \leq \gamma, & \text{noise is stationary;} \\ > \gamma, & \text{noise is nonstationary.} \end{cases} \qquad (3)$$

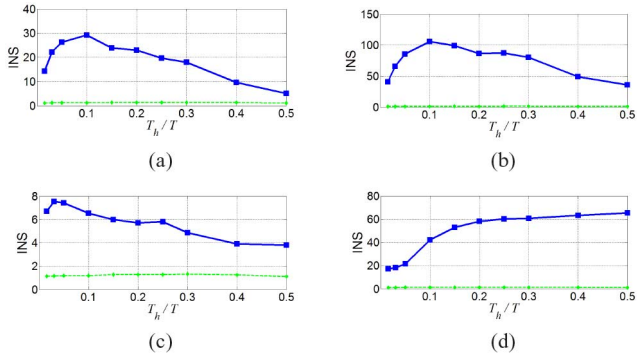

Fig. 2. The INS obtained for 3-seconds segments of the acoustic noise: (a) babble, (b) chainsaw, (c) factory (d) jackhammer. Dashed lines indicate the corresponding value for the threshold γ for the stationarity test.

The values of γ are also shown in the dashed lines of Fig. 2.

From these INS results it can be seen that the four noises are nonstationary relatively to all time scales. Chainsaw and jackhammer noises are here classified as highly nonstationary since they achieve INS values greater than 100 and 60, respectively. Babble noise presents INS results in a scale of 4 to 30 and thus, it is then designated as nonstationary. Factory noise is considered as moderately nonstationary since its INS values are lower than 8 for all time scales.

A. Speech Enhancement Baseline Techniques

Three speech enhancement techniques are here examined as baseline for the evaluation of the proposed solution: UMMSE, EMDF and EMDH.

1) UMMSE: The unbiased minimum mean-square error estimator [4] is adopted to track the noise spectrum. The authors combined the speech presence uncertainty to update the noise power spectrum every time frame by using a recursive procedure. UMMSE tracks nonstationary noises with shorter estimation delays than other estimators [3]. Moreover, a bias compensation factor is not required for the estimation.

2) EMDF: Firstly, the EMD-based filtering [7] decomposes the noisy speech signal into a set of intrinsic mode functions (IMF). Then, it identifies the number of IMFs that will be used in the speech signal reconstruction. The selection criteria is based on the IMF variances. The EMDF was defined to serve as a post-enhancement approach to the optimally-modified log spectral amplitude (OMLSA) [16] technique. In this work, the EMDF is directly applied to the noisy speech signals.

3) EMDH: This method employs the Hurst exponent (H) [11] as a criteria to identify the most corrupted IMFs on a frame-by-frame basis. The Hurst exponent expresses the time-dependence or scaling degree of a signal and is related to its spectral characteristics. The H values were used in [17] to compose a speech feature vector and successfully applied to speaker recognition. In [18], the Hurst exponent was also adopted for robust acoustic source localization. The selection criteria defined in [9] removes the IMFs whose Hurst exponent is above a given threshold. The remaining IMFs are then used to reconstruct the enhanced version of the speech signal. When compared to the UMMSE and EMDF techniques, the EMDH showed superior speech quality and intelligibility results for highly nonstationary noises [9].

Table I indicates the computational complexity which refers to the processing time required for each algorithm evaluated for 512 samples per frame. These values are normalized by the execution time of the proposed scheme (PRO). Note that PRO and UMMSE presented very low computational complexity when compared to the EMD-based algorithms.

TABLE I
NORMALIZED MEAN PROCESSING TIME

UMMSE	EMDF	EMDH	PRO
0.8	3.0	3.9	1.0

B. Speech Quality and Intelligibility Measures

Four objective measures are applied in the experiments. The segmental SNR (SegSNR) and the overall quality composite measure (OQCM) [19] are used to evaluate the proposed technique in terms of speech quality improvement while the short-time objective intelligibility measure (STOI) [20] and the coherence speech intelligibility index (CSII) [21] are adopted to examine the speech intelligibility.

1) Segmental SNR: The segmental SNR of a speech signal is defined as $\text{SegSNR} = \frac{10}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}} \log \frac{\sum_w |\mathbf{x}(w,q)|^2}{\sum_w |\eta(w,q)|^2}$, where $\mathbf{x}(w,q)$ and $\eta(w,q)$ are the STFT of the clean speech $x(k)$ and the noise $\eta(k)$, respectively, w is the frequency bin, q is the time frame index, $\mathcal{Q}$ is the set of frames of $x(k)$ with speech presence and $|\mathcal{Q}|$ its corresponding cardinality.

2) Overall Quality Composite Measure: The OQCM is a linear combination of three different objective measures: the weighted spectral slope (WSS), the log-likelihood ratio (LLR) and the perceptual evaluation of speech quality (PESQ), $\text{OQCM} = 1.594 + 0.805\text{PESQ} - 0.512\text{LLR} - 0.007\text{WSS}$. These coefficients were defined in [19] by using the multiple linear regression analysis to maximize the correlation between the OQCM values and the subjective speech quality results.

3) Coherence Speech Intelligibility Index: CSII is a spectral-based speech intelligibility measure [21] which is computed by multiplying coherence-based weights to the enhanced speech in the frequency domain. The signal is firstly split into segments using 30 ms Hamming windows. These segments are weighted by the magnitude-squared coherence between the clean and enhanced signals estimated across the entire signal. In this work, the predicted intelligibility scores are obtained by applying the following mapping function:

$$f(\text{CSII}) = \frac{100}{1+\exp(a\text{CSII}+b)}, \tag{4}$$

where $a = -10.09$ and $b = 4.65$.

4) Short-time Objective Intelligibility Measure: STOI [20] was proposed as a correlation-based method to evaluate the speech intelligibility degradation caused by speech enhancement solutions. A monotonic nonlinear mapping was applied to the STOI results to predict the percentage of correct words achieved in subjective listening tests. In this work, the predicted intelligibility scores are obtained by the mapping function (4) with $a = -13.45$ and $b = 9.36$.

C. Experiments Results

Fig. 3 depicts the SegSNR improvement achieved by the proposed and the baseline techniques with the four acoustic noises and SNR values. Note that PRO leads to the best SegSNR results

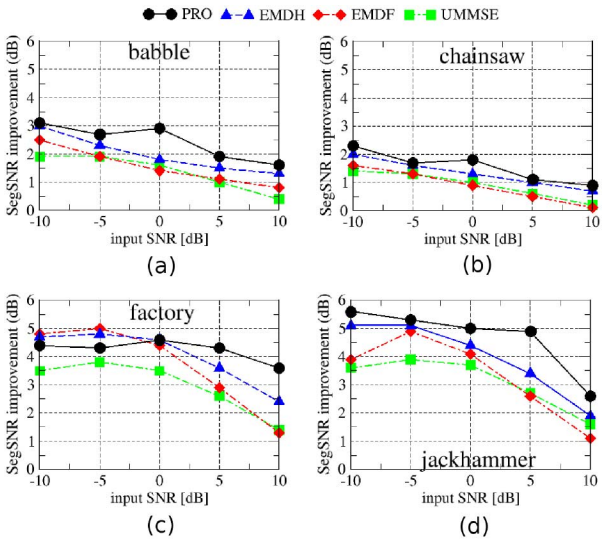

Fig. 3. The SegSNR improvement obtained with the proposed and the baseline techniques.

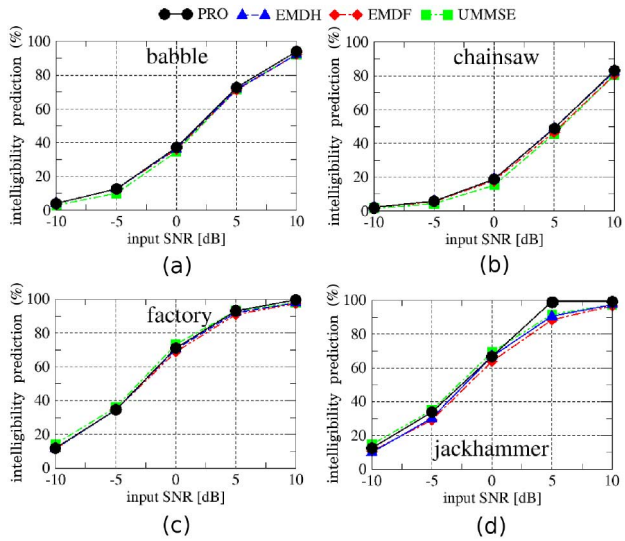

Fig. 4. Intelligibility rate prediction (%) obtained with CSII results.

TABLE II
OQCM IMPROVEMENT RESULTS

Noise	SNR	UMMSE	EMDF	EMDH	PRO
babble	10 dB	0.9	0.4	0.8	1.3
	5 dB	-0.1	0.5	1.0	1.3
	0 dB	-1.0	0.3	0.7	1.1
	-5 dB	-1.7	0.4	0.8	1.2
	-10 dB	-2.4	0.3	0.9	1.0
chainsaw	10 dB	0.9	0.3	1.4	1.7
	5 dB	0.0	0.9	1.6	1.9
	0 dB	0.9	1.2	1.7	2.2
	-5 dB	-1.5	1.5	2.3	2.1
	-10 dB	-2.7	1.3	2.4	2.1
factory	10 dB	3.4	0.7	0.9	1.6
	5 dB	2.8	1.0	1.5	1.9
	0 dB	2.3	1.9	2.3	2.7
	-5 dB	1.2	2.1	2.9	2.9
	-10 dB	0.1	1.8	2.3	2.2
jackhammer	10 dB	3.4	1.3	3.2	3.9
	5 dB	3.1	2.6	3.9	4.4
	0 dB	2.3	3.1	4.0	4.7
	-5 dB	0.9	3.2	3.8	3.6
	-10 dB	0.2	2.7	3.7	3.4

TABLE III
INTELLIGIBILITY RATE PREDICTION (%) OBTAINED WITH STOI

Noise	SNR	UMMSE	EMDF	EMDH	PRO
babble	10 dB	88.8	88.6	89.0	90.7
	5 dB	72.0	74.0	73.4	77.3
	0 dB	37.6	42.7	42.0	43.5
	-5 dB	9.2	12.9	12.5	11.9
	-10 dB	1.6	2.5	2.6	2.4
chainsaw	10 dB	85.7	84.4	88.2	90.0
	5 dB	57.0	59.0	61.8	66.7
	0 dB	19.3	24.4	25.1	31.2
	-5 dB	3.5	5.2	5.1	5.0
	-10 dB	1.1	1.5	1.5	1.3
factory	10 dB	93.5	89.8	93.1	94.4
	5 dB	87.5	85.7	87.2	89.8
	0 dB	73.5	74.7	72.1	70.4
	-5 dB	46.5	46.6	44.1	42.9
	-10 dB	17.3	14.7	14.3	13.5
jackhammer	10 dB	92.9	91.2	92.7	93.4
	5 dB	86.0	84.2	86.9	88.3
	0 dB	72.2	74.4	72.4	80.5
	-5 dB	44.1	46.6	42.4	43.7
	-10 dB	17.4	18.3	16.5	16.5

in almost all noisy conditions. For the highly nonstationary jackhammer noise and SNR of -10 dB, an improvement of 5.6 dB is obtained with the PRO method. The only scenarios where the PRO does not achieve the best results are for the factory noise with SNR < 0 dB. However, for this same noise source with SNR of 10 dB, PRO outperforms the baseline solutions in 1.2 dB.

The OQCM improvement scores attained with the PRO and the baseline techniques are presented in Table II. It can be seen that the proposed solution outperforms the other time domain techniques (EMDF and EMDH) for all the noise sources considering SNR ≥ 0 dB. When compared to the spectral-based UMMSE, PRO achieves the highest improvement for the nonstationary babble and the highly nonstationary chainsaw and jackhammer noises.

Fig. 4 presents the predicted intelligibility rates obtained with the CSII. The proposed solution outperforms all the other methods in 7% in average for the babble and chainsaw noises. Considering factory and jackhammer noises with SNR < 0 dB, the UMMSE leads to an average improvement of 5% over the other solutions. PRO achieves the best CSII results for SNR > 0 dB.

The predicted intelligibility rates computed with STOI are shown in Table III. PRO achieves the best STOI results for SNR ≥ 0 dB and for three noise sources: babble, chainsaw and jackhammer. Considering factory noise, PRO attains the highest intelligibility rates for SNR of 5 dB and 10 dB. For the highly nonstationary chainsaw noise, a STOI value of 90.0% is obtained with PRO, i.e., 4.3% higher than the UMMSE. In general, UMMSE leads to the best results with SNR < 0 dB.

IV. CONCLUSION

This letter introduced a novel time domain speech enhancement method for signals corrupted by nonstationary acoustic noise. Several experiments were conducted using acoustic noises with different INS and SNR values. The SegSNR, OQCM, STOI, and CSII objective measures demonstrated that the proposed technique outperforms the baseline approaches in terms of speech quality and intelligibility for all the acoustic noises.

REFERENCES

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, pp. 113–120, Apr. 1979.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, pp. 1109–1121, Dec. 1984.

[3] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, pp. 466–475, Sep. 2003.

[4] T. Gerkmann and R. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1383–1393, 2012.

[5] V.-K. Mai, D. Pastor, A. Aissa-El-Bey, and R. Le-Bidan, "Robust estimation of non-stationary noise power spectrum for speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 4, pp. 670–682, 2015.

[6] K. Manohar and P. Rao, "Speech enhancement in nonstationary noise environments using noise properties," *Speech Commun.*, vol. 48, pp. 96–109, Jan. 2006.

[7] N. Chatlani and J. Soraghan, "EMD-based filtering (EMDF) of low-frequency noise for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, pp. 1158–1166, May 2012.

[8] 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," in *Proc. Roy. Soc. London A: Math., Phys. Eng. Sci.*, Mar. 1998, vol. 454, pp. 903–995.

[9] L. Zão, R. Coelho, and P. Flandrin, "Speech enhancement with emd and hurst-based mode selection," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, pp. 899–911, May 2014.

[10] 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, FL, USA: CRC, 2015.

[11] E. Hurst, "Long-term storage capacity of reservoirs," *Amer. Soc. Civil Eng. Trans.*, pp. 770–799, Apr. 1951.

[12] D. Pastor and F.-X. Socheleau, "Robust estimation of noise standard deviation in presence of signals with unknown distributions and occurences," *IEEE Trans. Signal Process.*, vol. 60, no. 4, pp. 1545–1555, 2012.

[13] P. Borgnat, P. Flandrin, P. Honeine, C. Richard, and J. Xiao, "Testing stationarity with surrogates: A time-frequency approach," *IEEE Trans. Signal Process.*, vol. 58, pp. 3459–3470, Jul. 2010.

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

[15] 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 Commun.*, vol. 12, no. 3, pp. 247–251, 1993.

[16] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Process.*, vol. 81, no. 11, pp. 2403–2418, 2001.

[17] 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 Trans. Audio, Speech, Lang. Process.*, vol. 14, pp. 931–940, May 2006.

[18] E. Dranka and R. Coelho, "Robust maximum likelihood acoustic energy based source localization in correlated noisy sensing environments," *IEEE J. Sel. Topics Signal Process.*, vol. 9, pp. 259–267, Mar. 2015.

[19] Y. Hu and P. Loizou, "Evaluation of objective measures for speech enhancement," in *Proc. INTERSPEECH 2006*, Sep. 2006, pp. 1–4.

[20] C. Taal, R. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 19, pp. 2125–2136, Sep. 2011.

[21] J. Kates, "Coherence and the speech intelligibility index," *J. Acoust. Soc. Amer.*, vol. 4, pp. 2224–2237, Apr. 2005.