Speech Enhancement with EMD and Hurst-Based Mode Selection

L. Zão, *Member, IEEE*, R. Coelho, *Member, IEEE*, and P. Flandrin, *Fellow, IEEE*

***Abstract*—This paper presents a speech enhancement technique for signals corrupted by nonstationary acoustic noises. The proposed approach applies the empirical mode decomposition (EMD) to the noisy speech signal and obtains a set of intrinsic mode functions (IMF). The main contribution of the proposed procedure is the adoption of the Hurst exponent in the selection of IMFs to reconstruct the speech. This EMD and Hurst-based (EMDH) approach is evaluated in speech enhancement experiments considering environmental acoustic noises with different indices of nonstationarity. The results show that the EMDH improves the segmental signal-to-noise ratio and an overall quality composite measure, encompassing the perceptual evaluation of speech quality (PESQ). Moreover, the short-time objective intelligibility (STOI) measure reinforces the superior performance of EMDH. Finally, the EMDH is also examined in a speaker identification task in noisy conditions. The proposed technique leads to the highest speaker identification rates when compared to the baseline speech enhancement algorithms and also to a multicondition training procedure.**

***Index Terms*—Empirical mode decomposition, hurst exponent, index of nonstationarity, speaker identification, speech enhancement.**

I. Introduction

THE suppression of acoustic distortion in noisy speech signals is still an important research topic. The main issue of the speech enhancement techniques is concerned with the accurate estimation of the noise statistics, particularly, in real nonstationary environments. The classical estimators are based on voice activity detectors (VAD). The power spectrum of the noise components is then computed as a smoothed adaptation of its past values obtained during the speech pauses. These procedures show reasonable accuracy for stationary background noises but they cannot precisely estimate time-varying spectra. The difficulty in tracking nonstationary noises becomes more evident for long speech segments and low signal-to-noise ratio (SNR). The minimum statistics (MS) [1] and the improved minima controlled recursive averaging (IMCRA) [2] algorithms were proposed to deal with these situations. Thus, the estimation of the noise power spectrum is applied to each time frame even during speech activity. However, these approaches are inaccurate in tracking highly nonstationary noises [3]. Recent contributions, such as the unbiased minimum mean-square error (UMMSE) [4] algorithm, have been proposed to estimate the power spectrum of nonstationary noises with shorter delays.

In the literature, time-frequency (TF) analysis, e.g. wavelets, have also been adopted for speech enhancement. In such proposals [5], [6], the wavelet decomposition is applied to the noisy speech signal, and a decision criteria identifies the least corrupted components before the reconstruction of the enhanced version of the speech signal. Different from the power spectrum-based methods, the TF-based ones do not require explicit estimation of the noise statistics.

In the past few years, other TF speech enhancement solutions [7]–[9], based on the empirical mode decomposition (EMD) [10], have been introduced in the literature. The EMD is a nonlinear time-domain adaptive method for decomposing signals into a series of oscillatory intrinsic mode functions (IMF) and a residual. As opposed to the wavelet decomposition, the EMD does not require a set of basis functions to properly analyze the target signal. In fact, the IMFs obtained with the EMD depend only on the target data. Moreover, the EMD is not restricted to stationary signals. In [7], the EMD-based detrending (EMD-DT) technique was proposed to separate any kind of target signal from a corrupting slowly-varying trend. The EMD-based filtering (EMDF) was presented in [9] as a post-enhancement approach to remove residual low-frequency noise from previously enhanced speech signals. Although the EMDF showed promising objective quality results for speech corrupted with stationary noises, lower improvement was obtained with the nonstationary Babble noise (refer to [9]).

Speech enhancement techniques are generally evaluated in terms of their improvement in the speech quality. The segmental signal-to-noise ratio (SegSNR) and its frequency-domain version (the frequency-weighted SegSNR - fwSegSNR [11]) are examples of the commonly used speech objective quality measures. The spectral subtraction (SS) [12], the minimum mean-square error short-time spectral amplitude (MMSE-STSA) [13] and the optimally-modified log-spectral amplitude (OMLSA) [14] estimators are examples of approaches that achieve interesting objective quality improvement. However, a comparative study [15] of these noise-reduction algorithms showed that they are not capable of increasing the speech intelligibility. This situation becomes

Manuscript received October 30, 2013; revised January 29, 2014; accepted March 10, 2014. Date of publication March 19, 2014; date of current version March 31, 2014. This work was supported in part by the National Council for Scientific and Technological Development (CNPq) under the 304254/2012-6 research grant. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Roberto Togneri.

L. Zão is with the Graduate Program in Defense Engineering, Military Institute of Engineering (IME), Rio de Janeiro 22290-270, Brazil (e-mail: zao@ime.eb.br).

R. Coelho is with the Electrical Engineering Department, Military Institute of Engineering (IME), Rio de Janeiro 22290-270, Brazil (e-mail: coelho@ime.eb.br).

P. Flandrin is with the Physics Department (UMR 5672 CNRS), Ecole Normale Supérieure de Lyon, 69634 Lyon, France (e-mail: Patrick.Flandrin@enslyon.fr).

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/TASLP.2014.2312541

more challenging in nonstationary noisy scenarios due to the inaccurate noise statistics tracking [16].

This paper introduces a novel EMD-based speech enhancement technique in which the noise components of each IMF are identified and selected by its Hurst exponent [17] statistics. The resulting or least corrupted IMFs are used to reconstruct the enhanced version of the speech signal. In the proposed EMDH technique, the IMFs selection and the speech reconstruction are performed on a frame-by-frame basis. The EMDH is investigated considering both quality and intelligibility objective measures. It is shown that the proposed approach achieves speech intelligibility gain even in highly nonstationary noisy conditions. The EMDH technique is also evaluated as a post-enhancement approach to the OMLSA and the Wiener filtering algorithm [18] with the UMMSE noise estimator [4].

The EMDH evaluation experiments are conducted with speech signals corrupted with four real acoustic noises considering five different values of SNR. The experiments also include the computation of the index of nonstationarity (INS) [19] of the acoustic noises. Five baseline algorithms, namely SS, OMLSA, UMMSE, EMDF and EMD-DT, and four objective measures are adopted for the speech enhancement experiments. In terms of speech quality, the EMDH achieves the highest SegSNR and composite measure results for the highly nonstationary noises (e.g., Babble). Moreover, it outperforms the baseline EMDF and EMD-DT techniques for all the noise sources. The fwSegSNR and the short-time objective intelligibility (STOI) [20] measures are used to evaluate the intelligibility gain of the proposed and baseline methods. The best fwSegSNR improvement is obtained for the EMDH as a post-enhancement approach to the UMMSE. Regarding STOI, the EMDH outperforms the five baseline techniques.

The speech enhancement with the EMDH is also examined in speaker identification (SI) experiments conducted in noisy environments. The accuracy results show that the use of speech utterances processed with the EMDH substantially improves the overall SI performance in comparison to the noisy signals without use of the EMDH. Moreover, the adoption of EMDH leads to the best SI results when compared to the other speech enhancement techniques and also the use of a multicondition training procedure [21].

The remainder of this paper is organized as follows. Section II introduces the EMDH algorithm, including the basic concepts of the EMD and the definition of the Hurst exponent. Descriptions of the baseline speech enhancement techniques are presented in Section III. The objective measures used to evaluate the EMDH performance in terms of speech quality and intelligibility are briefly described in Section IV. The speech enhancement experiments are detailed in Section V. Then, the results obtained with the EMDH and the baseline approaches are presented and discussed. In Section VI, the basic concepts regarding the speaker identification task are introduced. The SI accuracy results obtained with the speech enhancement techniques are also presented in Section VI. Finally, Section VII concludes this work.

II. EMDH Speech Enhancement Technique

The first step of the proposed EMDH speech enhancement technique is to decompose the noisy speech signal into a set of IMFs using the EMD method. Then, the Hurst exponent is computed on a frame-by-frame basis from each of the resulting IMFs to determine which of them are mainly composed by noise. Finally, an enhanced version of the speech signal is reconstructed using the remaining IMFs.

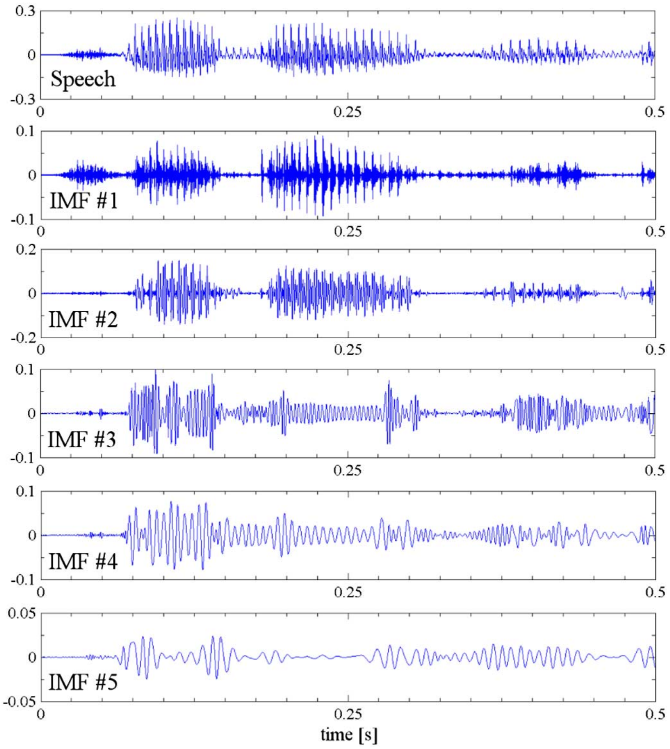


Fig. 1. The first five intrinsic mode functions obtained from the decomposition of a speech segment spoken by a male speaker.

In the literature, the wavelet decomposition has been widely used for time-frequency analysis. In this work, the EMD is adopted due to two main advantages over the wavelets-based approach. Firstly, the wavelet decomposition is based on a set of pre-defined basis functions, which does not necessarily fits well to all kinds of signals. Moreover, since it uses linear time-invariant filters, the wavelet decomposition is not adaptable to local or temporary variations in the input signal. On the other hand, the EMD analyzes the speech signal in an entirely adaptive way, and it is completely based on the local properties of the input signal. It makes the EMD suitable for nonstationary signal analysis and also assures the completeness of the signal reconstruction using the IMFs.

A. Empirical Mode Decomposition

The general idea of the EMD is to analyze a signal $x(t)$ between two consecutive extrema (minima or maxima), and defines a local high-frequency part, also called detail $d(t)$, and a local trend $a(t)$, such that $x(t) = d(t) + a(t)$. The first IMF is then composed of the local details, $d(t)$, obtained from all the consecutive extrema of $x(t)$. The high versus low-frequency separation procedure is iteratively repeated over the residual $a(t)$, leading to a new IMF and a new residual. Fig. 1 illustrates the first five IMFs obtained from decomposing a sample speech segment of 500 ms collected from the TIMIT [22] database.

The algorithm proposed in [10] for decomposing the input signal $x(t)$ can be summarized in the following steps:

1) Identify all extrema (local minima and maxima) of $x(t)$;
2) Obtain the upper ($e_{max}(t)$) and lower ($e_{min}(t)$) envelopes by interpolating[1] the local maxima and minima, respectively;
3) Compute the local trend as the average between the upper and lower envelopes, i.e., $a(t) = (e_{min}(t) + e_{max}(t))/2$;
4) Calculate the detail component as $d(t) = x(t) - a(t)$;
5) Iterate on the residual local trend $a(t)$.

The IMFs must have zero mean and all their local maxima and minima must be positive and negative, respectively[2]. If the detail component, obtained in step 4, does not follow these properties, steps 1 to 4 are repeated with $d(t)$ in place of $x(t)$. This process, called *sifting*, is repeated until the new $d(t)$ can be considered as an IMF. For the next IMF, the *sifting* process is applied on the residual $a(t) = x(t) - d(t)$.

From the EMD algorithm, it can be noticed that the total number of extrema is reduced from one IMF to the next. The waveform of each mode can be interpreted as a zero-mean amplitude and frequency modulated (AM-FM) signal. Note from Fig. 1 that the first IMF is composed of faster oscillations than the second, which in its turn has faster fluctuations than the third, and so on. It means that, at each time interval, the EMD applies a high-frequency versus low-frequency separation between IMFs. Thus, the first modes must present the high-frequency content of the signal. Moreover, as can also be noted from Fig. 1, the cutoff frequency between consecutive IMFs is time-varying and signal dependent.

Since the EMD algorithm can only be applied if there are at least two extrema in the last computed residual $a(t)$, any input signal $x(t)$ can be decomposed in a finite number of IMFs. If the m-th IMF is denoted as IMF_m and a total of M IMFs are extracted from $x(t)$, then

$$x(t) = \sum_{m=1}^{M} \text{IMF}_m(t) + r(t), \tag{1}$$

where $r(t)$ is the last residual obtained from the EMD algorithm.

In [23], it was shown that, when applied to fractional Gaussian noises (fGn), the EMD behaves like a dyadic filterbank with overlapping band-pass filters. In this analysis, the first IMF is interpreted as the output of a high-pass filter with a non-negligible content in its lower half-band. For the remaining modes, each IMF is roughly composed of the upper half-band part of the last residual $a(t)$ that results from the previous iteration.

B. EMDH: Hurst-based IMF Selection

The EMD algorithm states that, if a speech signal $x(t)$ is decomposed as in (1), its reconstruction using only a subset of the first N IMFs,

$$\tilde{x}(t) = \sum_{m=1}^{N} \text{IMF}_m(t), \text{ with } N < M, \tag{2}$$

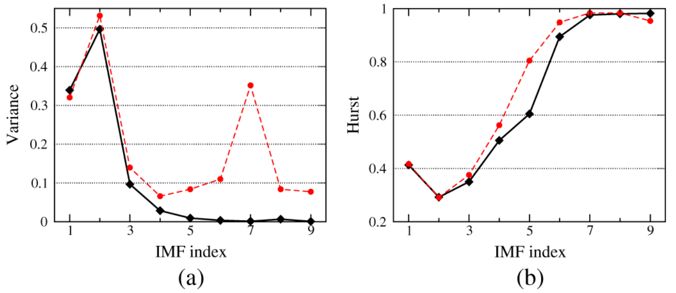


Fig. 2. The continuous lines indicate the values of (a) variance and (b) H of IMFs obtained from a clean speech utterance collected from TIMIT database. The dashed lines represent the corresponding values from the same speech segment corrupted by Factory noise with SNR of 0 dB.

would lead to the removal, at each time-frame, of the low-frequency components of $x(t)$. In [9], the authors showed that the energy content of a clean speech signal is mostly concentrated in the first four IMFs. Thus, they concluded that any value of $N \geq 4$ in (2) is enough for a good speech signal reconstruction.

The continuous line in Fig. 2(a) indicates the variance estimated from the samples of each IMF obtained from another speech utterance collected from the TIMIT database, i.e., $\sigma_m^2 = (1/T)\sum_t \text{IMF}_m^2(t), m = 1, 2, \ldots, 9$, where T is the total number of speech samples. It is noticeable that, in agreement with Fig. 1, there is an increase in the IMF energy (variance) from the first to the second IMF. Moreover, Fig. 2(a) also shows that the modes with the highest indices ($m > 4$) present lower energy values than the first ones. The dashed line in Fig. 2(a) represents the variance values obtained with the speech segment corrupted by a real Factory noise, extracted from NOISEX-92 database [24], with SNR of 0 dB. Note the sudden variance increase from IMFs 5 to 9, which is due to the low-frequency components of the corrupting noise.

The main issue of the proposed EMDH technique is the adoption of the Hurst exponent [17] to decide which IMFs should be selected for the speech signal reconstruction. Let the speech signal be represented by a stochastic process $x(t)$, with the normalized autocorrelation coefficient function (ACF, $\rho(k)$) defined by

$$\rho(k) = \frac{E\left[(x(t) - \mu_x)(x(t+k) - \mu_x)\right]}{E\left[(x(t) - \mu_x)^2\right]}, \tag{3}$$

where μ_x is the mean of $x(t)$ and k is the time lag. The ACF of a fractional Gaussian noise is given by [25]

$$\rho(k) = \frac{1}{2}\left(|k-1|^{2H} - 2|k|^{2H} + |k+1|^{2H}\right), \tag{4}$$

where $0 \leq H \leq 1$ is the Hurst exponent of $x(t)$. The H value is defined by the ACF decaying rate whose asymptotic behavior is

$$\rho(k) \sim H(2H-1)k^{2(H-1)}, \quad k \to \infty. \tag{5}$$

The Hurst exponent expresses the time-dependence or scaling degree of $x(t)$ and is related to its spectral characteristics. Within the whole range $]0, 1[$, the power spectral density $S_x(f)$ can be shown to be proportional to f^{1-2H} when $f \to 0$ [25]. For $H = 1/2$, $S_x(f)$ is constant over the whole frequency spectrum (e.g., white noise), whereas low frequencies are prominent in the case where $H > 1/2$, and in particular when

[1] Cubic splines are generally adopted to obtain the envelopes.

[2] These restrictions were defined since in the original EMD proposal [10] the IMFs are afterwards demodulated using the Hilbert transform.

$H \rightarrow 1$ ($1/f$ or pink noise). Due to such characteristics, the Hurst exponent was proposed in [26] to compose a speech feature vector and successfully applied to speaker recognition. In this work, the wavelet-based estimator [27] was adopted to obtain the H values of the IMFs on a frame-by-frame basis. The wavelet-based Hurst estimator can be described in three main steps as follows:

1) Wavelet decomposition: the discrete wavelet transform (DWT) is applied to successively decompose the input sequence of samples into approximation ($a_{\mathrm{w}}(j,n)$) and detail ($d_{\mathrm{w}}(j,n)$) coefficients[3], where j is the decomposition scale ($j = 1, 2, \ldots, J$) and n is the coefficient index of each scale.
2) Variance estimation: for each scale j, the variance $\sigma_j^2 = (1/N_j)\sum_n d_{\mathrm{w}}(j,n)^2$ is evaluated from the detail coefficients, where N_j is the number of available coefficients for each scale j. In [27], it is shown that $E[\sigma_j^2] = \mathcal{C}_H j^{2H-1}$, where $\mathcal{C}_H$ is a constant.
3) Hurst computation: a weighted linear regression is used to obtain the slope θ of the plot of $y_j = \log_2(\sigma_j^2)$ versus j. The Hurst exponent is estimated as $H = (1+\theta)/2$.

Fig. 2(b) illustrates the average values of H of different IMFs estimated from a TIMIT clean speech signal and the same corrupted by the Factory noise (Fig. 2(a)). The EMD is firstly used to decompose the speech signals. Then, the wavelet-based Hurst estimator is applied to each IMF (refer to Section II-B). The Hurst exponent is estimated from non-overlapping frames of 512 samples, which corresponds to 32 ms with sampling rate of 16 kHz, using the Daubechies filters [28] with 12 coefficients and the 3-12 scales. It can be seen that the first IMFs (e.g., 1-3), corresponding to the high frequency components, have $H < 1/2$. Moreover, for the highest IMF indices (e.g., 7-9) the H values are close to the unity, where the noise components are usually concentrated [29], [30]. This fact can also be observed in the speech signal corrupted with the Factory noise, where the low-frequency energy content ($H \approx 1$) is concentrated on the IMFs ≥ 7. It shows that the H exponent estimation enables the identification criteria to select the IMF low-frequency noise components.

C. EMDH Speech Signal Reconstruction

The EMDH algorithm starts with the decomposition of the input noisy speech into M modes according to (1). Windowed IMFs (w-IMF) are then obtained by splitting each mode into Q non-overlapping short-time frames,

$$\text{w-IMF}_{m,q}(t) = \begin{cases} \text{IMF}_m(t + qT_d), & t \in [0, T_d], \\ 0, & \text{elsewhere}, \end{cases} \tag{6}$$

where $q \in \{0, \ldots, Q-1\}$ is the frame index and T_d is the fixed time-duration of the frames. In a consecutive step, the wavelet decomposition is applied to all the windowed IMFs, w-IMF$_{m,q}(t)$, in order to estimate and store their Hurst exponent. Thus, a vector of Hurst values, $\mathbf{H}_q(m)$, with M components ($m = 1, \ldots, M$) are obtained for each frame index q. The next step is to determine, for each frame, the index N_q of the last windowed IMF whose value of H is below a given threshold, i.e., $\mathbf{H}_q(N_q) < H_{\mathrm{th}}$. If $\hat{x}(t)$ represents the enhanced speech signal, then each of its frames $\hat{x}_q(t)$ is reconstructed as

$$\hat{x}_q(t) = \sum_{m=1}^{N_q} \text{w-IMF}_{m,q}(t), q = 0, \ldots, Q-1, \tag{7}$$

and $\hat{x}(t)$ is finally given by

$$\hat{x}(t) = \sum_{q=0}^{Q-1} x_q(t - qT_d). \tag{8}$$

In the proposed EMDH, the IMF selection is exclusively based on the Hurst exponent estimated from short-time segments. This frame-by-frame analysis avoids that sudden changes in the power spectrum of nonstationary noises affect the IMF selection of the entire speech signal.

To avoid discontinuities, the following procedure is applied in the signal reconstruction. Suppose that the speech frame q is reconstructed with a smaller number of w-IMFs than the next. Thus, there is at least one index m' such that w-IMF$_{m',q}(t)$ is included in the reconstruction of frame q, but w-IMF$_{m',q+1}(t)$ is not in frame $q+1$. Then, the samples of the half-right part of w-IMF$_{m',q}(t)$ are multiplied by the samples of the half-right part of the Hanning window whose size equals the frame duration. Therefore, the value of the last sample of w-IMF$_{m',q}(t)$ turns to zero and the continuity of the reconstructed speech signal is preserved. The analogous procedure is adopted when any IMF is used in the reconstruction of frame $q+1$ and not of frame q.

III. Speech Enhancement Baseline Techniques

This Section briefly describes the five baseline speech enhancement techniques adopted in this work. The SS, OMLSA and UMMSE apply the short-time Fourier transform (STFT) to firstly obtain an estimate of the noise power spectrum. Following, the identified noise components are subtracted or compensated from the STFT of the noisy signal to improve the speech quality.

A. Spectral Subtraction

Let $y(t)$ be a speech utterance corrupted by an additive noise $\eta(t)$. Thus, it can be written $y(t) = x(t) + \eta(t)$, where $x(t)$ represents the clean speech signal. By applying the STFT to the above relation, it can be written

$$Y(\kappa, \tau) = X(\kappa, \tau) + \mathcal{N}(\kappa, \tau), \tag{9}$$

where κ and τ are the frequency bin and the time frame indices, respectively.

The first step of SS [12], [31] is to estimate the noise power spectrum $|\hat{\mathcal{N}}(\kappa, \tau)|^2$ using the classical VAD-based approach. Then, the clean speech power spectrum is estimated as [31]

$$|\hat{X}|^2 = \max\left\{|Y|^2 - \alpha|\hat{\mathcal{N}}|^2, \beta|\hat{\mathcal{N}}|^2\right\} \tag{10}$$

In (10), the spectral floor parameter (β) and the time-varying oversubtraction factor (α) are set as in [31]. The spectrum of the enhanced signal is then estimated using the phase of the noisy speech signal, and the enhanced speech signal $\hat{x}(t)$ is

[3] The subscript /w/ is used to discriminate the detail ($d(t)$) and trend ($a(t)$) components of EMD, from the detail ($d_{\mathrm{w}}(j,n)$) and approximation ($a_{\mathrm{w}}(j,n)$) coefficients of the wavelet decomposition.

finally reconstructed by overlapping and adding its inverse Fourier transform.

B. OMLSA

The second baseline technique adopted in this work applies the IMCRA [2] to obtain an estimate of the noise power spectrum. Then, the OMLSA [14] is used to reconstruct the enhanced version of the clean speech. The IMCRA noise estimator is composed of two iterations. Firstly, a VAD is defined based on the minimum noisy speech power spectrum values obtained from a set of past frames. In a second stage, this VAD is used to determine the speech presence probability $p(\kappa,\tau) \in [0,1]$ for each frequency bin and each time frame. The noise power spectrum estimation $|\bar{\mathcal{N}}(\kappa,\tau)|^2$ is recursively given by

$$|\bar{\mathcal{N}}(\kappa,\tau+1)|^2 = \delta_\eta|\bar{\mathcal{N}}(\kappa,\tau)|^2 + (1-\delta_\eta)|Y(\kappa,\tau)|^2, \quad (11)$$

where $\delta_\eta(\kappa,\tau)$ is a time-varying smoothing parameter that depends on $p(\kappa,\tau)$.

After the noise spectrum estimation, the OMLSA method reconstructs the enhanced speech signal $\hat{x}(t)$ by minimizing the mean-square error of the log-spectral amplitude. The gain function that leads to the spectral amplitude $|\hat{X}(\kappa,\tau)|$ of the optimally reconstructed speech is defined in [14] as

$$G_{\mathrm{OMLSA}}(\kappa,\tau) = \{G_{\mathrm{LSA}}(\kappa,\tau)\}^{p(\kappa,\tau)} G_{\min}^{1-p(\kappa,\tau)}, \quad (12)$$

where $G_{\mathrm{LSA}}(\kappa,\tau)$ is a function of the *a priori* SNR (refer to [14]), and the minimum value $G_{\min}$ is defined by a subjective criteria. All the parameters used in the OMLSA and IMCRA implementation, including the bias compensation factor for the noise estimation, are the same as those adopted in [2], [14].

C. UMMSE

In the third speech enhancement baseline procedure, the unbiased minimum mean-square error (UMMSE) noise power estimation [4] is adopted to track the noise spectrum. In this proposal, the authors combined speech presence uncertainty to the estimator originally proposed in [32], and found that the estimation of the noise power spectrum $|\hat{\mathcal{N}}(\kappa,\tau)|^2$ can be updated every time frame via the recursive smoothing

$$|\hat{\mathcal{N}}(\kappa,\tau)|^2 = \alpha_p|\hat{\mathcal{N}}(\kappa,\tau-1)|^2 + (1-\alpha_p)E\left(|\mathcal{N}|^2|Y\right) \quad (13)$$

where α_p is a smoothing factor and the noise periodogram estimate $E(|\mathcal{N}|^2|Y)$ depends on the speech presence and absence probabilities and on the noise power spectrum estimated from the last frame. The main issue of adopting the UMMSE is that, unlike IMCRA, it does not require a minimum search within a given number of past frames. It leads to shorter delays in the noise estimation. Besides, UMMSE does not require a bias compensation factor.

Following the procedure in [4], the UMMSE noise estimator is followed by the speech enhancement algorithm proposed in [18]. The Wiener filtering gain is based on the estimation of the *a priori* SNR, which is obtained with the decision-directed approach proposed in [13]. The UMMSE approach was implemented in C++ and validated by reproducing the results of noise power estimation obtained with the MATLAB code provided by the authors [4].

D. EMDF

The main issue of the EMD-based filtering is to identify the number N of IMFs that will be used in the speech signal reconstruction, according to (2). In [9], the authors showed that, for clean speech, the variance of $\mathrm{IMF}_m(t)$ decreases for the highest IMF index m. In Section II-A (refer to Fig. 2(a)), it was also demonstrated that, in the case of low-frequency noise corruption, variance peaks would appear at IMFs with high indices. Thus, a selection criteria is defined based on the IMF variances. In these situations, the IMF index N is determined by the minimum variance value that occurs prior to the identified peak. The EMDF algorithm can be described as follows:

1) Decompose the target speech signal $x(t)$ using EMD, as in (1);
2) Compute the variances from the samples of each IMF, i.e., $V(m) = \sigma_m^2 = (1/T)\sum_{t=1}^{T} \mathrm{IMF}_m^2(t)$, where T is the total number of speech samples;
3) Identify the first variance peak m_p such that $V(m_p-1) < V(m_p)$ and $V(m_p+1) < V(m_p)$, with $m_p > 4$;
4) Find the index m_t of the minimum variance value that occurs prior to m_p, which means that $V(m_t) < V(m_t+1)$, $V(m_t) < V(m_t-1)$ and $m_t < m_p$;
5) Reconstruct the speech signal with N IMFs, according to (2), with $N = m_t$.

In the example of speech corrupted with Factory noise shown in Fig. 2(a), the identified indices are $m_p = 7$ and $m_t = 4$. In [9], the EMDF was proposed as a post-enhancement approach to the OMLSA technique. In this work, it is also directly applied to the noisy speech signals.

E. EMD-DT

The EMD-based detrending and denoising method was proposed in [7] as a simple way of splitting a target signal from superimposed slow oscillations. In the EMD-DT, the last IMF of index used in the signal reconstruction, is defined by the standardized means of the IMFs. The idea is to remove the IMFs whose empirical standardized mean significantly departs from zero. For this purpose, the standardized mean of each IMF is computed as

$$\mathrm{StdMean}(m) = \frac{\frac{1}{T}\sum\limits_{t=1}^{T}\mathrm{IMF}_m(t)}{\sqrt{\frac{1}{T}\sum\limits_{t=1}^{T}\mathrm{IMF}_m^2(t)}}, \quad (14)$$

and the EMD-DT algorithm searches for the first IMF, with index $N+1$ ($N \geq 4$), for which $\mathrm{StdMean}(N+1)$ is compared to the root mean square of the standardized mean of the first four modes multiplied by a threshold ζ, i.e.

$$|\mathrm{StdMean}(N+1)| > \zeta\sqrt{\frac{1}{4}\sum_{m=1}^{4}\mathrm{StdMean}^2(m)}. \quad (15)$$

Finally, the enhanced version of the target signal is reconstructed as in (2). In this work, the threshold in (15) is empirically set to $\zeta = 2$.

IV. Speech Objective Quality and Intelligibility Measures

This Section introduces the objective quality and intelligibility measures adopted for the evaluation of the EMDH technique. While the SegSNR and the composite measure are used to evaluate the speech quality, the fwSegSNR [11] and the STOI [20] are considered for intelligibility.

A. Segmental SNR

The time-domain segmental SNR is used to measure speech quality and it is defined as

$$\mathrm{SegSNR} = \frac{10}{Q}\sum_{\tau=0}^{Q-1}\log\frac{\sum_{t_s=\tau T_{\mathrm{sh}}}^{\tau T_{\mathrm{sh}}+T_d-1} x^2(t_s)}{\sum_{t_s=\tau T_{\mathrm{sh}}}^{\tau T_{\mathrm{sh}}+T_d-1}\left[x(t_s)-\hat{x}(t_s)\right]^2}, \quad (16)$$

where T_d is the frame length (in samples), T_{sh} is the frame shift, Q is the total number of frames, and $x(t_s)$ and $\hat{x}(t_s)$ are the discrete-time representations of the clean and enhanced speech signals, respectively. In this work, the SegSNR values are obtained with frame size of 32 ms with 75% overlapping corresponding to the values $T_d = 512$ and $T_{\mathrm{sh}} = 128$ samples with 16 kHz sampling rate. For the SegSNR computation, the SNR of each frame is limited between -10 dB and 35 dB [11]. In Section V, the results are presented in terms of SegSNR improvement, which is here defined as the SegSNR from the enhanced speech subtracted from the values obtained from the noisy signals. The same definition is adopted for the composite and fwSegSNR improvement.

B. Overall Composite Quality Measure

In [33], the authors evaluated the correlation between five objective measures and three subjective rating scores: signal distortion, background noise distortion and overall quality. Then, in order to achieve higher correlation with the subjective scores, three composite measures were proposed as the linear combination of the existing objective measures. For the overall speech quality, the composite measure was defined as [33]

$$C_{\mathrm{ovl}} = 1.594 + 0.805\mathrm{PESQ} - 0.512\mathrm{LLR} - 0.007\mathrm{WSS} \quad (17)$$

where PESQ is the perceptual evaluation of speech quality, LLR is the log-likelihood ratio and WSS is the weighted spectral slope distance. In this work, the overall composite measure (17) is computed considering the wideband version of PESQ, as defined by the ITU-T recommendation P.862.2.

C. Frequency-Weighted SegSNR

The adoption of the frequency-weighted SegSNR is motivated by the experiments results described in [16], which demonstrated that the fwSegSNR is highly correlated to the subjective speech intelligibility. For the fwSegSNR computation, the spectra of the clean ($|X(j,\tau)|$) and enhanced ($|\hat{X}(j,\tau)|$) speech signals are obtained by dividing their entire bandwidth into $K = 25$ frequency bands using Gaussian-shaped filters. Then, the fwSegSNR is computed as

$$\mathrm{fwSegSNR} = \frac{10}{Q}\sum_{\tau=0}^{Q-1}\frac{\sum_{j=1}^{K} W(j,\tau)\log\frac{|X(j,\tau)|^2}{\left(|X(j,\tau)|-|\hat{X}(j,\tau)|\right)^2}}{\sum_{j=1}^{K} W(j,\tau)}, \quad (18)$$

where τ and j are the frame and frequency band indices, respectively. As proposed in [11], the signal-dependent weighting function is defined by $W(j,\tau) = |X(j,\tau)|^{(0.2)}$. The SNR values computed at each frame and each frequency band are also limited to range of $[-10, 35]$.

D. Short-Time Intelligibility Objective Measure

The short-time objective intelligibility measure [20] was proposed as a correlation-based method to evaluate the speech intelligibility degradation caused by the speech enhancement procedures. It was shown in [20] that STOI has high and very close correlation to subjective intelligibility rates obtained with speech signals enhanced by noise-reduction algorithms. Following the procedure presented in [20], the clean and the noisy versions of the speech signal are divided into short-time frames and grouped in 15 one-third octave bands. For each frame τ and each band j, the intermediate intelligibility measure, $\mathrm{STOI}_{(j,\tau)}$, is defined as the correlation coefficient between the temporal envelope vectors obtained from the clean and the noisy speech signals. Finally, the STOI measure is given by averaging the intermediate values over the 15 one-third octave bands and all Q speech frames.

In [20], a monotonic nonlinear mapping was applied to the STOI results to predict the percentage of correct words achieved in subjective listening tests with native people. The results showed good precision considering the enhanced speech signals from two evaluated databases. In this work, the predicted intelligibility scores are obtained by applying a mapping function to the STOI results,

$$f(\mathrm{STOI}) = \frac{100}{1+\exp(a\mathrm{STOI}+b)}, \quad (19)$$

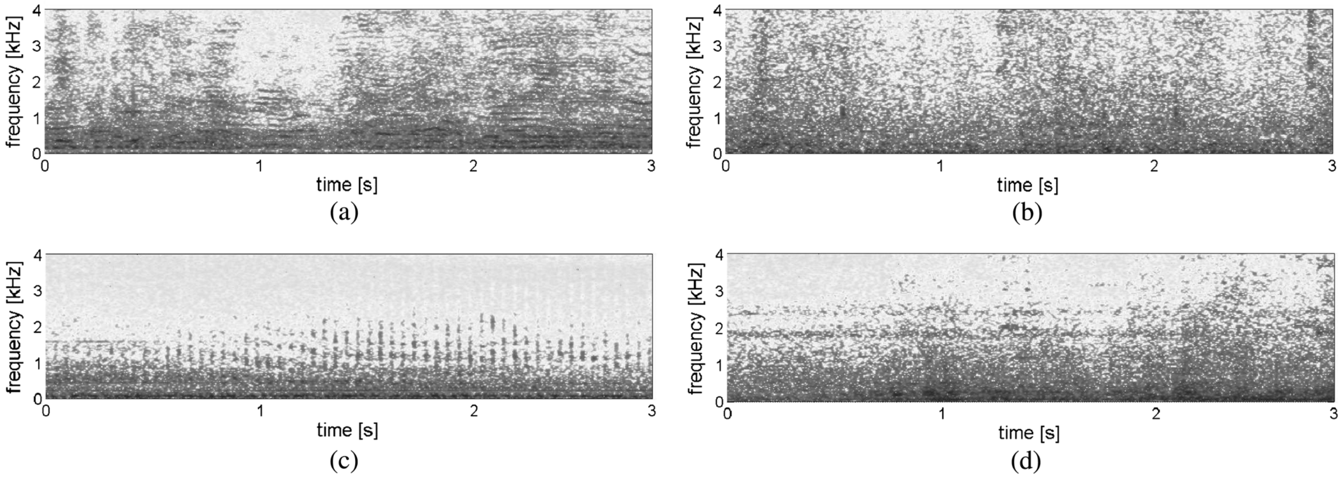
with $a = -17.4906$ and $b = 9.6921$. The values of a and b in (19) are the same as those found in [20]. However, it is important to mention that their exact values are not crucial since $f(\mathrm{STOI})$ is a monotonically increasing function for any $a < 0$. It means that, a higher value of $f(\mathrm{STOI})$ also implies a higher value for STOI. On the other hand, the evaluation of intelligibility rate prediction ($f(\mathrm{STOI})$) instead of the absolute STOI value provides a more practical way to examine the intelligibility of the speech enhancement procedures.

V. Speech Enhancement Evaluation Experiments

In this Section, the objective measures presented in Section IV are firstly used to measure the EMDH performance in terms of speech quality (composite, SegSNR) and intelligibility (fwSegSNR, STOI). The STOI is followed by the mapping function in (19) to predict the speech intelligibility scores.

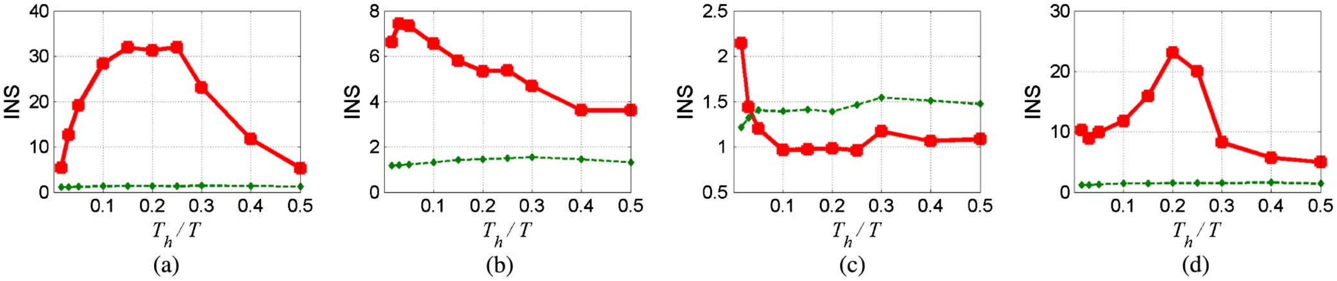
For the speech enhancement experiments, a subset of 24 speakers (16 male and 8 female) is randomly selected from the TIMIT speech database [22]. It leads to a total of 240 speech segments, 10 per speaker, with sampling rate of 16 kHz and average time duration of 3 seconds.

Four acoustic noises (Babble, Factory, Helicopter and Train) are used to corrupt the speech signals considering five SNR values: -10 dB, -5 dB, 0 dB, 5 dB and 10 dB. The noises



(a) (b) (c) (d)

Fig. 3. Spectrograms for 3-seconds segments of the acoustic noises: (a) Babble, (b) Factory, (c) Helicopter, and (d) Train.



(a) (b) (c) (d)

Fig. 4. The INS values obtained for 3-seconds segments of the acoustic noises: (a) Babble, (b) Factory, (c) Helicopter, and (d) Train. Dashed lines indicate the corresponding values for the threshold γ for the stationarity tests.

are collected from the NOISEX-92 [24] and the Freesound.org[4] databases: Babble and Factory from the former and Helicopter and Train from the latter. The time-varying spectral behavior of the selected noises are shown in Fig. 3. Note that Babble, Factory and Train noises present spectral components fluctuating over the entire voice frequency band (0-4 kHz). On the other hand, the spectrogram of the Helicopter noise is characterized by constantly repetitive energy impulses mainly at frequency values lower than 2.5 kHz.

In the EMDH algorithm presented in Section II-C, the value of the threshold $H_{\rm th}$ is crucial to determine the portion of the low-frequency noise that is removed from each speech frame. If $H_{\rm th} \approx 1$, the low-frequency noise components will be removed from the signal. The suppression of the low-frequency components is due to the spectral characteristics of the acoustic noises adopted in this work (refer to Fig. 3). Moreover, such kind of low-frequency spectrum is widely found in real acoustic noises [29]. Although any other value can be adopted for the $H_{\rm th}$, in the following experiments it is set to $H_{\rm th} = 0.9$ in order to remove the noise components without distorting the speech signal.

The index of nonstationarity [19] is here adopted as a time-frequency approach to objectively examine the nonstationarity of each acoustic noise. The INS values computed from segments of the four noises, are depicted in the continuous lines of Fig. 4. The time scale T_h/T is the ratio of the length adopted in the short-time spectral analysis (T_h), and the total time duration ($T = 3$ seconds) of the noises sample sequences. The noises are considered as nonstationary whenever their INS values are above the threshold γ [19], which are shown in the dashed lines of Fig. 4.

The INS results indicate that the Babble, Factory and Train noises are nonstationary for all time scales. Babble and Train noises achieve INS values greater than 30 and 20, respectively. Since for most of the time scales their INS values are substantially greater than the stationarity threshold defined as $\text{INS} > 10\gamma$, these noises are considered as highly nonstationary. As it can be seen, Factory noise is also nonstationary but as its INS values are lower than 8, it can be identified as moderately nonstationary. On the other hand, the Helicopter noise achieves very low INS values (≈ 1) and, in general, below the defined stationarity threshold. Thus, it is here considered as stationary noise. It is interesting to mention that the short-time impulses shown in Fig. 3(c) are captured by the INS of the Helicopter noise for the shortest time scale.

A. Speech Quality Evaluation

The proposed EMDH speech enhancement technique is firstly evaluated in terms of the segmental SNR. The SegSNR improvement (in dB) obtained with the proposed and the baseline approaches are depicted in Fig. 5. The top curves represent the SegSNR gain for the highly nonstationary noises, and those in the bottom are obtained from the noises with lower INS. Note that, in comparison to the other EMD-based techniques, the EMDH presents the highest improvement for most of the noise conditions. When compared to the SS, OMLSA and UMMSE, the EMDH also achieves the best performance for

[4] http://www.freesound.org.

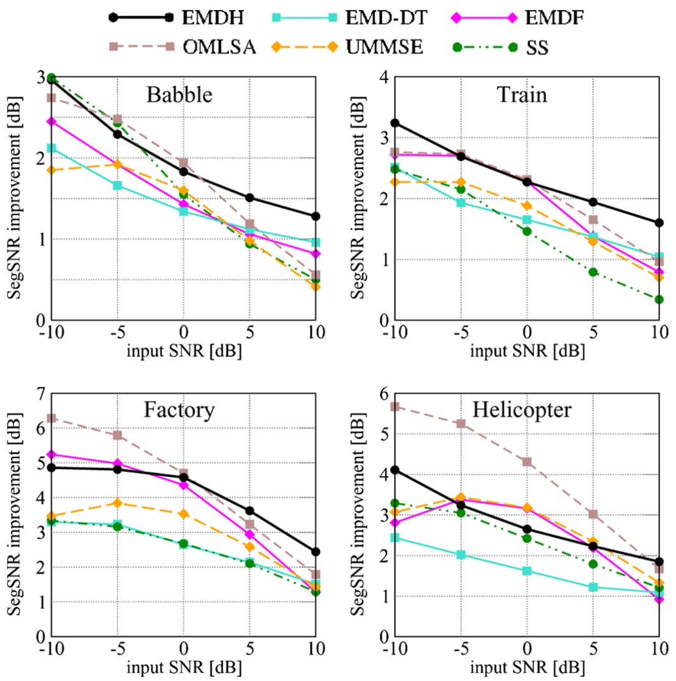


Fig. 5. The SegSNR improvement (dB) obtained with the EMDH and the baseline techniques.

most of the input SNR values considering the highly nonstationary noises. The OMLSA leads to the highest SegSNR gain for the other noise sources. However, even for such noises the proposed EMDH achieves the greatest SegSNR improvement for the highest SNR values, i.e., Factory with SNR $\geq$ 5 dB and Helicopter with SNR of 10 dB.

Fig. 6 shows the improvement in the SegSNR result obtained with the EMDF and EMDH applied as post-enhancement to the OMLSA and UMMSE techniques. In this scenario, the best results are achieved with the OMLSA followed by the EMDH. Note that the adoption of the EMDH improves the SegSNR results for all the four noise sources, mainly for the highly nonstationary ones. In comparison to the OMLSA results (Fig. 5), the greatest contribution of the proposed technique is for Babble noise. For this noise source with SNR of -10 dB, the SegSNR gain is increased from 2.7 dB with OMLSA to 4.1 dB with OMLSA followed by EMDH, corresponding to a difference of 1.4 dB.

As a complement to the segmental SNR, the overall quality composite measure is also evaluated for the speech enhancement approaches. The improvement obtained with the EMDH and the baseline techniques are depicted in Fig. 7. Note that the EMDH outperforms the other EMD-based techniques for almost all the input SNR values. The only conditions for which the EMDF and the EMDH present similar performance are the Train and Helicopter noises with SNR of 0 dB. When compared to the STFT-based algorithms, the EMDH also achieves the highest improvement for the highly nonstationary noises, i.e., Babble and Train. For the other two noises, the OMLSA leads to the best performance for SNR $\geq$ 0 dB, while the EMDH obtains the highest C_{ovl} gain for SNR $<$ 0 dB.

It can also be noted from Fig. 7 that, unlike the EMDH, the STFT-based techniques decrease the overall quality composite

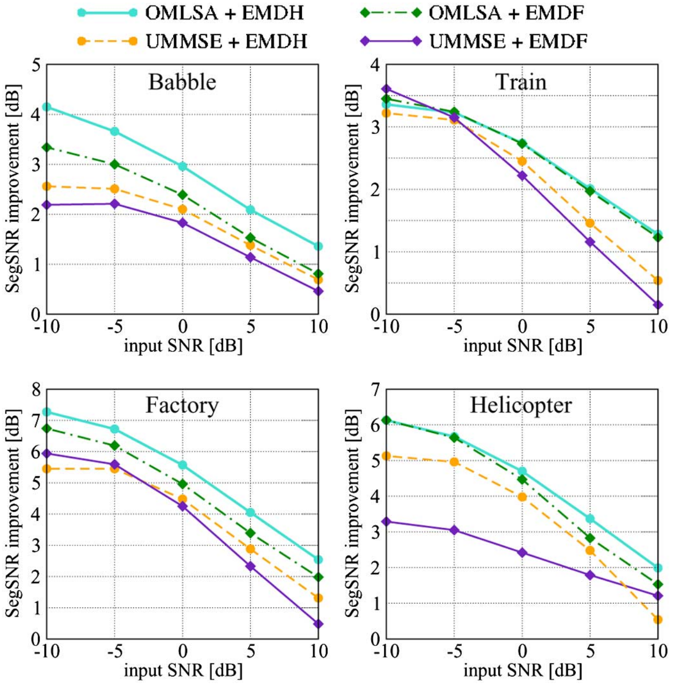


Fig. 6. The SegSNR improvement (dB) with the EMDH and EMDF as post-enhancement to the OMLSA and UMMSE techniques.

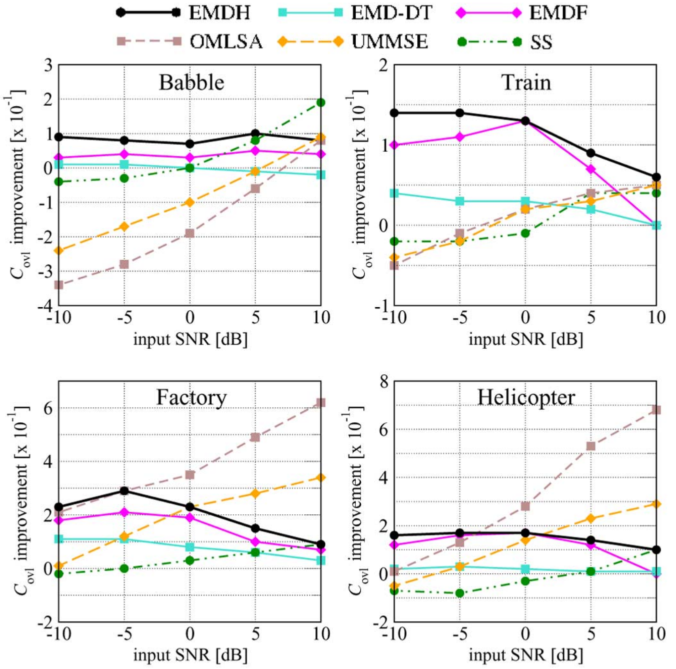


Fig. 7. The overall quality composite measure improvement obtained with the EMDH and the baseline techniques.

measure results for the highly nonstationary noises with SNR $<$ 0 dB. The poor performance in such conditions can be explained by the inaccurate estimation of the time-varying power spectra of such acoustic noises. Even the UMMSE approach, which adopts a noise tracking method with shorter delays, is not able to accurately suppress the noise components from the speech signal. Since the EMDH technique does not require the estimation of the noise components, it seems to be a good solution for situations with highly nonstationary noises.

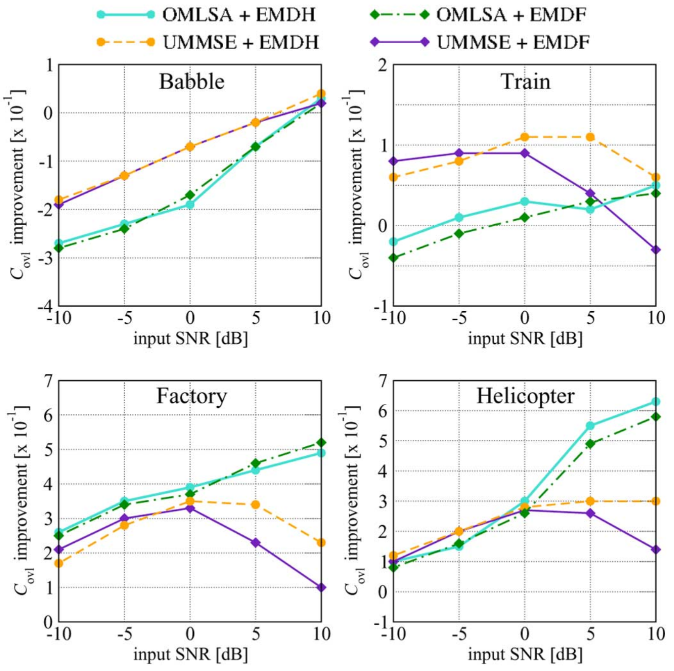


Fig. 8. The overall quality composite measure improvement with the EMDH and EMDF as post-enhancement to the OMLSA and UMMSE techniques.

The composite results considering the post-enhancement procedures are shown in Fig. 8. Here, the EMDH also outperforms the EMDF for most of the noise conditions. Again, the speech enhancement approaches that adopts the OMLSA technique achieve the highest improvement for the noises with the lowest INS values, i.e., Factory and Helicopter. For the Train noise, the UMMSE followed by the EMDH obtains the best results for the SNR $\geq$ 0 dB. Finally, both the UMMSE + EMDH and UMMSE + EMDF approaches present similar performance for Babble noise.

B. Speech Intelligibility Evaluation

The fwSegSNR gain obtained from the proposed and baseline speech enhancement procedures are illustrated in Fig. 9. Once more, considering the EMD-based techniques, the EMDH achieves the best results for most of the input SNR values for the three nonstationary noise sources. For the Factory noise, for example, the EMDH achieves a fwSegSNR improvement of 2.0 dB for SNR of 5 dB. When compared to the STFT-based approaches, the EMDH also leads to the best fwSegSNR results for the Babble noise. Different from the segmental SNR results shown in Fig. 5, the UMMSE outperforms the OMLSA algorithm for Factory noise. The OMLSA technique leads to the highest improvement for Helicopter and Train noises.

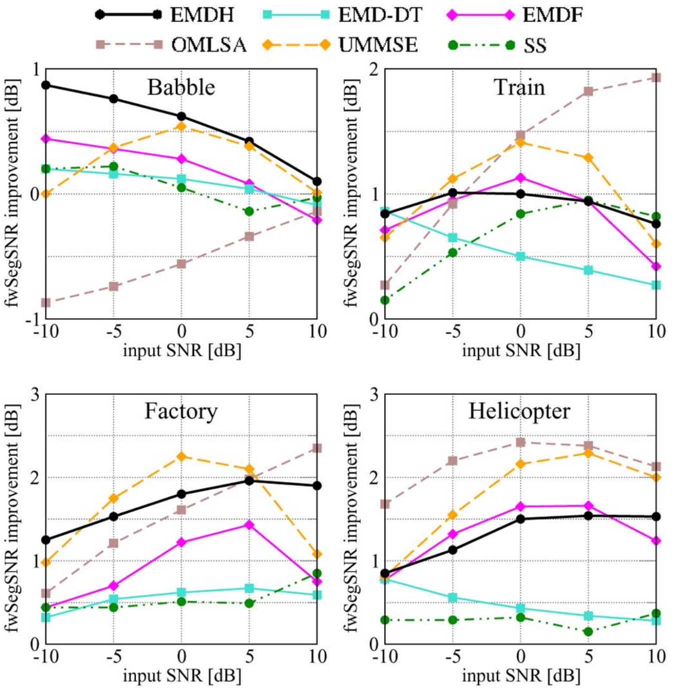


Fig. 9. The fwSegSNR improvement (dB) obtained with the EMDH and the baseline techniques.

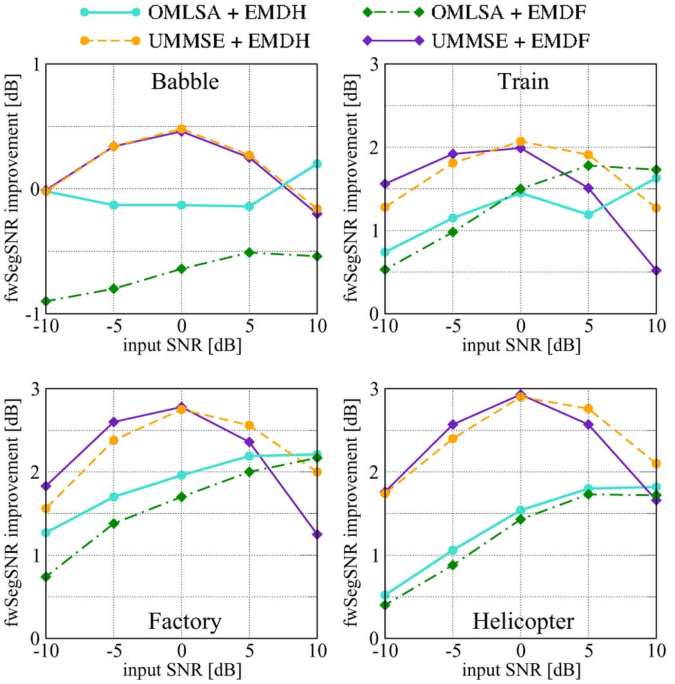


Fig. 10. The fwSegSNR improvement (dB) with the EMDH and EMDF as post-enhancement to the OMLSA and UMMSE techniques.

Fig. 10 shows the fwSegSNR improvement obtained with EMDF and EMDH as post-enhancement techniques. It is worth to mention that the EMDH also outperforms the EMDF for almost all the noise conditions. Different from the objective quality measures (Figs. 6 and 8), the best performance is here achieved with the speech enhancement procedures that adopt the UMMSE technique. The UMMSE + EMDH leads to the highest fwSegSNR improvement for most of the input SNR values considering all the acoustic noise sources. From Figs. 9 and 10, it can be observed that the OMLSA technique provides negative gain for Babble noise, also in post-enhancement scenarios. Similar results were found in [15] and it means that, although speech enhancement techniques improve speech quality, they can also degrade speech intelligibility. The STOI is here adopted as a complementary measure to examine the EMDH performance in terms of speech intelligibility.

C. STOI Prediction

In Table I, the intelligibility prediction rates are obtained from the STOI of the processed speech signals followed by the mapping function in (19). It can be noted that the highest and lowest

TABLE I
INTELLIGIBILITY RATE PREDICTION (%) OBTAINED WITH STOI RESULTS FOLLOWED BY THE MAPPING FUNCTION IN (19)

Noise	SNR	SS	OMLSA	UMMSE	EMDF	EMD-DT	EMDH
Babble	10	99.60	99.62	99.53	99.56	99.50	99.58
	5	97.64	98.21	97.97	98.27	97.98	98.25
	0	78.72	86.60	87.15	89.90	88.82	89.73
	-5	31.45	41.49	44.60	54.29	51.59	53.32
	-10	6.27	7.73	10.22	14.76	14.31	15.50
	Aver.	62.74	66.73	67.89	71.35	70.44	71.28
Train	10	99.62	99.60	99.53	99.55	99.53	99.58
	5	98.84	98.83	98.58	98.65	98.65	98.78
	0	94.69	95.89	95.09	95.44	95.06	95.72
	-5	68.26	82.30	81.09	80.97	80.47	81.72
	-10	21.48	38.47	41.70	41.60	40.98	42.40
	Aver.	76.58	83.02	83.20	83.24	82.94	83.64
Factory	10	99.77	99.83	99.72	99.64	99.78	99.77
	5	99.04	99.47	99.30	99.41	99.39	99.52
	0	92.72	97.99	97.65	98.42	97.84	98.27
	-5	67.86	91.15	89.89	91.58	89.40	91.42
	-10	31.51	56.19	59.74	58.76	56.25	59.10
	Aver.	78.18	88.93	89.26	89.56	88.53	89.62
Helicopter	10	99.66	99.78	99.76	99.57	99.69	99.74
	5	98.07	99.32	99.32	99.03	98.97	99.18
	0	87.54	97.43	97.36	96.78	95.87	96.76
	-5	52.77	87.00	87.85	84.14	81.41	83.86
	-10	19.28	48.09	54.73	44.80	42.82	45.82
	Aver.	71.46	86.33	87.80	84.87	83.75	85.07
Overall result		72.24	81.25	81.81	82.26	81.42	82.40

STOI results for all the speech enhancement procedures are obtained for Factory and Babble noises, respectively. All noise sources lead to high intelligibility scores ($f(\text{STOI}) > 97\%$) for SNR ≥ 5 dB. However, large differences in STOI prediction is found with the lowest SNR values. It can be seen that the proposed EMDH technique outperforms the other EMD-based approaches for three acoustic noises: Factory, Helicopter and Train. The proposed technique also leads to the best overall STOI results: 82.40%.

Note from Table I that, when compared to the SS, OMLSA and UMMSE, the EMDH achieves the best average results for the three nonstationary noises. For instance, the predicted intelligibility score with Babble noise is improved from 67.89% with UMMSE to 71.28% with EMDH. The overall STOI result for OMLSA and UMMSE are 81.25% and 81.81%, respectively.

The results in Figs. 7–10 and in Table I emphasize that the proposed EMDH technique increases the time and frequency-domain SegSNR and also achieves the highest overall STOI prediction scores. In the next Section, the EMDH is evaluated in speaker identification experiments conducted in noisy environments. For this purpose, the EMDH is used as a pre-processing step for the SI system and its performance is compared to other speech enhancement procedures adopted in this work (SS, OMLSA, UMMSE, EMDF and EMD-DT).

VI. ISSUES ON SPEAKER IDENTIFICATION

Speech enhancement solutions have been examined to provide robustness to speaker identification systems [34]–[36]. In this work, SI experiments are conducted to evaluate the contribution of the proposed EMDH and the baseline speech enhancement techniques on improving the accuracy of SI in nonstationary noisy conditions.

A speaker identification system is generally composed of a training and a test phase [37]. During the training phase, the system extracts the sets of speech features and generates the speakers models. In the test phase, the speech features are obtained from the test speech utterances and compared to the speakers models. The main goal of the SI task is to identify to which of the enrolled speakers the test utterance belongs to. In the literature, SI systems based on the Mel-frequency cepstral coefficients (MFCC) [38] features and the Gaussian mixture speaker model (GMM) [37] are widely used due to their high recognition accuracies for clean speech [39]. However, their performance can be severely degraded when the test speech signals are corrupted by acoustic noises, i.e., a noisy mismatch condition between training and test phases [40]. As shown in [21], significant improvement can be achieved by using a colored-noise-based multicondition training (Colored-MT). In the experiments here described, the EMDH technique is applied to the test speech utterances to provide noise robustness to the SI system. This means that the MFCC feature vectors are extracted from the enhanced versions of the speech signals.

1) MFCC Extraction: After the acquisition and pre-processing, the speech signal is divided into overlapping short-time frames. The fast Fourier transform (FFT) is applied to each speech frame, and its spectral envelope is then obtained using Mel-scaled bandpass filters [38]. Frequencies in the Mel scale (f_{Mel}) are related to frequencies in the linear scale (f_{Hz}) as $f_{\text{Mel}} = 1127\log(1 + \frac{f_{\text{Hz}}}{700})$. The Mel scale is usually applied in speaker identification due to its good representation of the human auditory system. Considering J the number of filters in the Mel-frequency filterbank [38] and E_j the log-energy output of the jth filter, the MFCC coefficients are calculated as $\text{MFCC}_i = \sum_{j=1}^{J} E_j \cos[i(j - \frac{1}{2})\frac{\pi}{K}], i = 1, 2, \ldots, D$, where D is the number of cepstrum coefficients. As commonly adopted in the literature [37] [40], $J = 26$ filters in the Mel-scaled filterbank are used for the MFCC extraction.

2) GMM: The GMM (λ_S) of a speaker S is defined as a linear combination of Gaussian components, $p(\vec{x}|\lambda_S) = \sum_{n=1}^{\mathcal{M}} p_n b_n(\vec{x})$, where $\vec{x}$ is a D-dimensional speech feature vector, p_n are the mixture weights, with $\sum_{n=1}^{\mathcal{M}} p_n = 1$, and $b_n(\vec{x})$ are the Gaussian densities with mean vectors $\vec{\mu}_n$ and covariance matrices K_n. Thus, the GMM of speaker S can be parametrized by $\lambda_S = \{p_n, \vec{\mu}_n, K_n | n = 1, \ldots, \mathcal{M}\}$.

During training, the parameters of λ_S are estimated as to maximize the likelihood function $p(\mathbf{X}|\lambda_S) = \prod_{t=1}^{Q} p(\vec{x}_t|\lambda_S)$, where the speech feature matrix $\mathbf{X}$ is composed of Q feature vectors $\vec{x}_t$ extracted from each frame of the training utterance available for speaker S. For the tests, the decision rule of the SI task is based on the maximum log-likelihood criteria [37]. It means that the identified speaker $\hat{S}$ is the one that maximizes the sum $\hat{S} = \arg\max_S \sum_{t=1}^{Q} \log p(\vec{x}_t|\lambda_S)$.

3) Colored-MT: The multicondition training was proposed due to the mismatch between the training and test phases caused by acoustic noise corruption. The idea is to improve the SI system robustness by artificially corrupting the training utterances. For this purpose, the authors in [40] used white noise as-

TABLE II
SPEAKER IDENTIFICATION ACCURACIES (%) OBTAINED WITH AND WITHOUT SPEECH ENHANCEMENT

Noise	SNR	NSE	SS	OMLSA	UMMSE	EMDF	EMD-DT	EMDH
Babble	20	96.7	67.3	49.7	72.0	94.4	95.2	97.3
	15	89.3	59.2	42.0	62.5	87.5	91.1	91.7
	10	59.2	44.6	26.2	44.1	63.7	65.2	76.2
	5	27.4	25.3	14.0	22.6	31.0	26.5	33.0
	0	7.4	9.8	9.5	9.8	8.6	8.6	8.3
	Aver.	56.0	41.3	28.3	42.2	57.0	57.3	61.3
Train	20	86.0	62.8	45.5	14.9	92.6	93.5	97.3
	15	71.4	55.1	36.3	26.8	86.0	84.2	94.1
	10	49.1	42.3	28.6	39.6	69.6	61.3	78.3
	5	32.4	27.1	22.6	51.8	44.9	38.4	46.1
	0	13.4	13.7	14.0	64.9	21.1	17.9	20.5
	Aver.	50.5	40.2	29.4	39.6	62.9	59.1	67.3
Factory	20	78.0	62.5	52.1	67.3	91.7	92.9	97.0
	15	60.1	49.1	51.2	59.5	84.5	80.1	94.9
	10	43.8	33.3	40.8	46.1	69.6	56.3	85.4
	5	21.1	22.6	28.9	31.9	38.4	31.3	53.0
	0	8.0	12.5	17.3	17.3	15.8	10.4	16.7
	Aver.	42.2	36.0	38.0	44.4	60.0	54.2	69.4
Helicopter	20	74.7	61.6	49.1	65.8	89.9	87.2	95.5
	15	56.3	45.2	42.0	52.7	76.5	66.7	83.6
	10	33.6	25.9	33.9	35.1	50.9	38.7	58.3
	5	15.2	14.6	22.0	19.4	20.8	13.4	23.8
	0	7.1	6.0	9.8	9.2	6.9	5.1	3.9
	Aver.	37.4	30.7	31.4	36.4	49.0	42.2	53.0
Overall		46.5	37.0	31.8	40.7	57.2	53.2	62.8

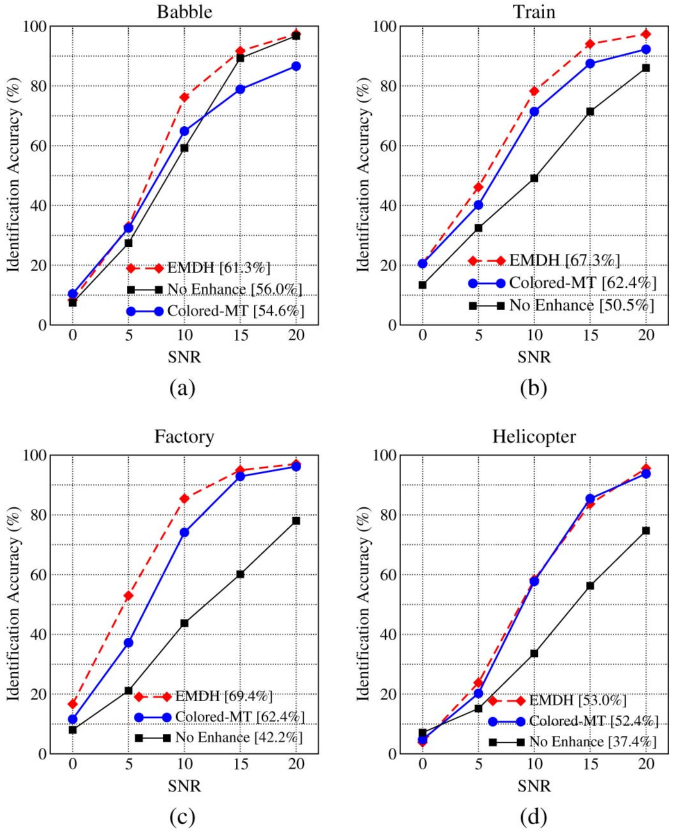


Fig. 11. Speaker identification accuracies obtained with EMDH, Colored-MT and without any enhancement, considering four different acoustic noise sources: (a) Babble, (b) Train, (c) Factory, and (d) Helicopter. The average results are indicated in the legends.

suming no information concerning the acoustic noises is available. In [21] it was proposed the use of artificially generated colored-spectra noises to obtain higher SI results. The motivation is that colored spectra have been measured in several acoustic noises [29]. Finally, the maximum log-likelihood criteria is adapted to consider all the speaker models obtained from the multicondition data.

4) Experiments and Results: The SI experiments are conducted with a subset of 168 speakers from TIMIT database. From each of the 10 utterances available per speaker, eight are concatenated and used to train the speaker models and the other two are separated for tests. Each of the $168 \times 2 = 336$ test utterances are then corrupted with the four noises considering SNR values of 0 dB, 5 dB, 10 dB, 15 dB and 20 dB. Therefore, a total of $20 \times 336 = 6720$ tests are conducted, leading to an accuracy precision of 0.015 considering a confidence degree of 95%. The speech feature vectors are composed by $D = 12$ coefficients, extracted from frames of 32 ms with 50% overlapping. $\mathcal{M} = 32$ Gaussian densities are used for each speaker model.

The third column of Table II presents the identification accuracies, in %, obtained from the SI experiments conducted with no speech enhancement (NSE). As a reference, the identification rate considering clean test utterances is 98.9%. Note that the identification rate varies from 96.7% with Babble noise and SNR of 20 dB to 7.1% for Helicopter noise and SNR of 0 dB. In average, the SI accuracy ranges from 56.0% to 37.4% for these same noise sources.

The SI results obtained with the speech enhancement techniques are also shown in Table II. It shows that the EMDH outperforms the baseline approaches for all the noise sources. The average accuracy increases from 46.5% with noisy speech to 62.8% with utterances processed by the EMDH technique, corresponding to 16.3 percentage point (p.p.) gain. Considering the different noise conditions, the best improvement is obtained for the Factory noise with SNR of 10 dB, from 43.8% to 85.4%, corresponding to 41.6 p.p. difference.

Regarding the other EMD-based approaches, the proposed technique improves the overall SI accuracy in 5.6 p.p. and 9.6 p.p. in comparison to the EMDF and the EMD-DT, respectively. It is important to notice that, although outperformed by EMDH, the EMDF and EMD-DT also improve the SI performance for all the noise sources. On the other hand, the SS and OMLSA techniques degrade the SI accuracies for the four noises, while the UMMSE improves the average identification rate only for Factory noise. It is interesting to mention that, the EMD-based techniques which achieved the best STOI results (refer to Table I), also presented the best identification rates. This indicates that future speech enhancement proposals should consider the SI aspects to improve their intelligibility gain.

Fig. 11 compares the speaker identification results obtained with the proposed EMDH technique (dashed lines) to those obtained with the Colored-MT (thick continuous lines) for each noise source. The corresponding average identification scores are shown in the legends. The accuracies obtained with noisy speech (i.e., no speech enhancement) are also depicted with thin continuous lines in Fig. 11. For the Colored-MT, $R = 3$ artificial acoustic noises are generated according to [30] with Gaussian distribution and power spectral density proportional to $1/f^{\delta}$. As in [21], the colored-spectra are defined by $\delta = 0$ (white noise), $\delta = 1$ (pink noise) and $\delta = 2$ (red or brown noise). The colored noises are used to corrupt the training utterances with SNR of 15 dB, since it led to the best overall results in preliminary test. The results in Fig. 11 show that the EMDH technique leads to the best average identification results for the four noise sources. The overall result obtained with Colored-MT is 58.0%, i.e., 4.8 p.p. lower than EMDH.

VII. Conclusion

This paper has introduced a new speech enhancement technique based on EMD and on a Hurst-based IMF selection criteria. The Hurst exponent statistics is adopted to identify and select those IMFs that are most affected by the noise components. The speech signal is finally reconstructed considering the least corrupted IMFs. Several experiments were conducted using four different acoustic noises, three of them nonstationary. The EMDH performance was compared to five baseline speech enhancement algorithms, and it was also evaluated as a post-enhancement approach. The proposed technique improved four objective measures that are highly correlated with speech quality and intelligibility. Moreover, the EMDH outperformed the baseline EMDF and EMD-DT approaches for most of the noise conditions. The superior performance of the proposed speech enhancement technique was also verified in the speaker identification experiments conducted in noisy environments.

Acknowledgment

The authors would like to thank the anonymous reviewers for their important remarks and suggestions that helped to improve this work.

References

[1] R. Martin, “Noise power spectral density estimation based on optimal smoothing and minimum statistics,” *IEEE Trans. Speech Audio Process.*, 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. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, Sep. 2003.

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

[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, May 2012.

[5] D. Donoho and I. Johnstone, “Threshold selection for wavelet shrinkage of noisy data,” in *Proc. 16th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC’94)*, Nov. 1994, vol. 1, pp. A24–A25.

[6] D. Donoho, “De-noising by soft-thresholding,” *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 613–627, May 1995.

[7] P. Flandrin, P. Gonçalves, and G. Rilling, “Detrending and denoising with empirical mode decompositions,” in *Proc. Eur. Signal Process. Conf. (EUSIPCO ’04)*, Sep. 2004, pp. 1581–1584.

[8] K. Khaldi, A. Boudraa, A. Bouchikhi, and M. Alouane, “Speech enhancement via EMD,” *EURASIP J. Adv. Signal Process.*, vol. 2008, no. 1, p. 873204, May 2008.

[9] N. Chatlani and J. Soraghan, “EMD-based filtering (emdf) of low-frequency noise for speech enhancement,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1158–1166, May 2012.

[10] 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. R. Soc. London. Ser. A: Math., Phys., Eng. Sci.*, Mar. 1998, vol. 454, no. 1971, pp. 903–995.

[11] Y. Hu and P. Loizou, “Evaluation of objective quality measures for speech enhancement,” *IEEE Trans. Audio, Speech. Lang. Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

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

[13] 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, no. 6, pp. 1109–1121, Dec. 1984.

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

[15] P. Loizou and Y. Hu, “A comparative intelligibility study of single-microphone noise reduction algorithms,” *J. Acoust. Soc. Amer.*, vol. 22, no. 3, pp. 1777–1786, Sep. 2007.

[16] J. Ma, Y. Hu, and P. Loizou, “Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions,” *J. Acoust. Soc. Amer.*, vol. 125, no. 5, pp. 3387–3405, May 2009.

[17] E. Hurst, “Long-term storage capacity of reservoirs,” *Amer. Soc. Civil Eng. Trans.*, no. 11, pp. 770–799, Apr. 1951.

[18] P. Scalart and J. Filho, “Speech enhancement based on a priori signal to noise estimation,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Dec. 1996, vol. 32, no. 6, pp. 629–632.

[19] 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, no. 7, pp. 3459–3470, Jul. 2010.

[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, no. 7, pp. 2125–2136, Sep. 2011.

[21] L. Zão and R. Coelho, “Colored noise based multicondition training technique for robust speaker identification,” *IEEE Signal Process. Lett.*, vol. 18, no. 11, pp. 675–678, Nov. 2011.

[22] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, “TIMIT acoustic-phonetic continuous speech corpus,” in *Linguist. Data Consortium*, Philadelphia, PA, USA.

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

[24] 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, Jul. 1993.

[25] B. Mandelbrot and J. Van Ness, “Fractional brownian motions, fractional noises and applications,” *SIAM Rev.*, vol. 10, no. 4, pp. 422–437, Oct. 1968.

[26] 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, no. 3, pp. 931–940, May 2006.

[27] D. Veitch and P. Abry, “A wavelet-based joint estimator of the parameters of long-range dependence,” *IEEE Trans. Inf. Theory*, vol. 45, no. 3, pp. 878–897, Apr. 1999.

[28] I. Daubechies, *Ten lectures on wavelets*. Philadelphia, PA, USA: Soc. Ind. and Appl. Math., 1992.

[29] R. Voss and J. Clarke, “$1/f$ noise in music: Music from $1/f$ noise,” *J. Acoust. Soc. Amer.*, vol. 63, no. 1, pp. 258–263, Jan. 1978.

[30] L. Zão and R. Coelho, “Generation of coloured acoustic noise samples with non-gaussian distribution,” *IET Signal Process.*, vol. 6, no. 7, pp. 684–688, Sep. 2012.

[31] M. Berouti, R. Schwartz, and J. Makhoul, “Enhancement of speech corrupted by acoustic noise,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP ’79)*, Apr. 1979, vol. 4, pp. 208–211.

[32] R. Hendriks, R. Heusdens, and J. Jensen, “MMSE based noise PSD tracking with low complexity,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP ’10)*, Mar. 2010, pp. 4266–4269.

[33] Y. Hu and P. Loizou, “Evaluation of objective measures for speech enhancement,” in *Proc. INTERSPEECH ’06*, Sep. 2006, pp. 1–4.

[34] J. Ortega-Garcia and J. Gonzalez-Rodriguez, “Overview of speech enhancement techniques for automatic speaker recognition,” in *Proc. 4th Int. Conf. Spoken Lang. (ICSLP ’96)*, Oct. 1996, vol. 2, pp. 929–932.

[35] S. Sadjadi and J. Hansen, “Assessment of single-channel speech enhancement techniques for speaker identification under mismatched conditions,” in *Proc. INTERSPEECH*, Sep. 2010, pp. 2138–2141.

[36] C. Maina and J. Walsh, “Joint speech enhancement and speaker identification using approximate bayesian inference,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 6, pp. 1517–1529, Aug. 2011.

[37] D. Reynolds and R. Rose, "Robust text independent speaker identification using gaussian mixture speaker models," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 72–82, Jan. 1995.

[38] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 4, pp. 357–366, Aug. 1980.

[39] D. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," *Speech Commun.*, vol. 17, pp. 91–108, Aug. 1995.

[40] J. Ming, T. Hazen, J. Glass, and D. Reynolds, "Robust speaker recognition in noisy conditions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1711–1723, Jul. 2007.

Leonardo Zão obtained the Ph.D. degree from the Military Institute of Engineering (IME) of Rio de Janeiro in 2013. From the same Institute, he received the M.Sc. and B.Sc. degrees in Electrical Engineering in 2010 and 2005, respectively. Since 2013, he has worked at the Laboratory of Acoustic Signal Processing (LASP/IME) as a Research Assistant. His current research mainly focuses on speaker recognition, speech enhancement, speech emotion classification and acoustic signal processing.

Rosângela Fernandes Coelho received the Ph.D. degree from the Ecole Nationale Supérieure des Télécommunications (ENST-Paris) in 1995 and the M.Sc. degree from the Pontifícia Universidade Católica of Rio de Janeiro (PUC-Rio) in 1991, both in Electrical Engineering.

She joined the Military Institute of Engineering (IME) of Rio de Janeiro, in 2002, where she is Associate Professor at the Electrical Engineering Department. Prof. Coelho founded and heads the Laboratory of Acoustic Signal Processing. In 2003, she received the University Research Program research grant from CISCO/USA. She also served as editorial board member of the IEEE Communications Surveys and Tutorials from 1999-2007. Since 2008, she has been responsible for the International Scientific Collaboration IME-ParisTech that includes 10 french engineering schools. Prof. Coelho was President-Adjoint of the Brazilian Telecommunications Society from 2008-2010 and she is member of the Signal Processing Society. In 2011, Prof. Coelho received the USPTO patent of an automatic speaker recognition method based on a new speech feature and speaker classifier. Her main research interests include acoustic signal processing, speech enhancement and intelligibility, speech and speaker recognition, time-frequency analysis, acoustic emotion detection and classification, acoustic speech features, acoustic signal and noise representation and generation, non-stationary noise, and statistical signal processing.

Patrick Flandrin (M'85–SM'01–F'02) received the engineer degree from ICPI Lyon, France, in 1978, and the Doct.-Ing. and "Docteur d'État" degrees from INP Grenoble, France, in 1982 and 1987, respectively. He joined CNRS in 1982, where he is currently Research Director. Since 1991, he has been with the "Signals, Systems and Physics" Group, within the Physics Department at École Normale Supérieure de Lyon, France. In 1998, he spent one semester in Cambridge, UK, as an invited long-term resident of the Isaac Newton Institute for Mathematical Sciences and, from 2002 to 2005, he has been Director of the CNRS national cooperative structure "GdR ISIS." He is currently President of GRETSI, the French Association for Signal and Image Processing. His research interests include mainly nonstationary signal processing (with emphasis on time-frequency and time-scale methods), self-similar stochastic processes and complex systems. He published many research papers in those areas and he is the author of the book Temps-Fréquence (Paris: Hermès, 1993 and 1998), translated into English as *Time-Frequency/Time-Scale Analysis* (San Diego: Academic Press, 1999). He has been a guest co-editor of the Special Issues "Wavelets and Signal Processing" of the IEEE TRANS. SIGNAL PROCESSING in 1993 and "Time-Frequency Analysis and Applications" of the *IEEE Signal Processing Magazine* in 2013. Past Associate Editor for the IEEE TRANS. SIGNAL PROCESSING (1993-1996 and 2008-2011) and former member of the "Signal Processing Theory and Methods" Technical Committee of the IEEE Signal Processing Society (1993-2004), he is currently on the Editorial Board of the *IEEE Signal Processing Magazine*.

Dr. Flandrin was awarded the Philip Morris Scientific Prize in Mathematics (1991), the SPIE Wavelet Pioneer Award (2001), the Prix Michel Monpetit from the French Academy of Sciences (2001) and the Silver Medal from CNRS (2010). He is a Fellow of IEEE (2002), of EURASIP (2009), and a member of the French Academy of sciences since 2010.