Impulsive Noise Detection for Speech Enhancement in HHT Domain

C. Medina, R. Coelho, *Senior Member, IEEE*, and L. Zão, *Member, IEEE*

Abstract—**This paper introduces a novel single channel speech enhancement method in the time domain to mitigate the effects of acoustic impulsive noises. The ensemble empirical mode decomposition is applied to analyze the noisy speech signal. The estimation and selection of noise components is based on the impulsiveness index of decomposition modes. An adaptive threshold is proposed to define the criterion to select the noise components. The proposed method is evaluated in speech enhancement experiments considering four acoustic noises with different impulsiveness indices and non-stationarity degrees under various signal-to-noise ratios. Four speech enhancement algorithms are adopted as baseline in the evaluation analysis considering spectral and time domains. Seven objective measures are adopted to compare the proposed and baseline approaches in terms of speech quality and intelligibility. Results show that the proposed solution outperforms the competing algorithms for most of the noisy scenarios. The novel method shows particularly interesting performance when speech signals are corrupted by highly impulsive acoustic noises.**

Index Terms—**Speech enhancement, impulsive noises, Hilbert-Huang transform, non-stationary acoustic noises.**

I. Introduction

IMPULSIVE background noisy condition may cause severe impact on the accuracy of acoustic classification systems and applications. Impulsive noises (falling objects, industrial machinery, slamming doors) are encountered in real environments. Impulsive noise may also cause severe impairment on the human auditory system [1]. They are commonly characterized by almost instantaneous sharp sounds with high acoustic energy and wide spectral bandwidth. Impulsive sample sequences are generally defined in the literature by heavy-tail distributions tailored by its impulsiveness degree. Due to this impulsive nature, a key element of the research area includes the accurate estimation of noise components especially from real acoustic noisy signals.

In the literature, many studies have been dedicated to mitigate the effect of real acoustic noise in different domains [2]–[4]. Most popular speech enhancement techniques apply the short-time Fourier transform (STFT) to process the noisy signal in the frequency domain [5], [6]. Since some impulsive disturbances may affect only a few samples of a single speech frame, frequency-domain approaches are less suitable to deal with impulsive acoustic noises [7]. Additionally, the required use of the original noisy phase for the reconstruction of the enhanced signal is also a limitation of most of these techniques [8].

In recent years, speech enhancement solutions have also been proposed in the time domain [3], [4], [7], [9], [10]. In [3], for example, interesting speech quality results are achieved in different noise scenarios after noise samples are subtracted directly from the speech signal. The acoustic noise statistics are estimated without any assumption regarding the speech samples distribution. Some other approaches [4], [11] apply harmonic models to represent voiced speech segments as series of sinusoids whose frequencies are given as multiples of the speech fundamental frequency. The main limitation relies on the fact that such harmonic models cannot be used in unvoiced speech regions. Other time-domain speech enhancement methods are based on the Hilbert-Huang Transform (HHT). Particularly, the Empirical Mode Decomposition (EMD) [12] or one of its variations have been adopted to analyze the noisy speech signal [9], [10], [13], [14] and other recent tasks [15]. EMD-based approaches have achieved promising speech quality and intelligibility improvement in non-stationary noisy scenarios [9], [10]. Impulsive noises may be considered as a different kind of non-stationary sources.

This work introduces the HHT-based method to enhance single channel speech signals corrupted by impulsive acoustic noise. The proposed HHT-α solution applies the ensemble EMD (EEMD) [16] to decompose a target noisy signal into a series of intrinsic mode functions (IMF). The noise components of each IMF are then identified and selected based on the impulsiveness index α [17]. The speech signal is reconstructed excluding frames that are mainly composed by noise. The selection criterion proposed here aims to remove most of the noise components without distorting the speech dominant segments of the signal. By exploiting the intrinsic nature of the impulsive noise, i.e., few speech samples can be highly corrupted by the acoustic noise, the proposed solution enables interesting quality and intelligibility improvement. Furthermore, the impulsive masking components detection also promotes a natural and interesting speech source separation. In the HHT-α, no assumption is considered for the speech and noise distribution. It also avoids the use of voice activity detector (VAD) or voiced/unvoiced separation. The selection threshold value is empirically determined using speech signals corrupted by two highly impulsive acoustic

Manuscript received July 17, 2020; revised November 28, 2020 and June 11, 2021; accepted June 17, 2021. Date of publication June 30, 2021; date of current version July 14, 2021. The work of R. Coelho was supported in part by the National Council for Scientific and Technological Development (CNPq) under Grant 308155/2019-0 and in part by the Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ) under Grant 26/203.075/2016. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Tom Bäckström. *(Corresponding author: Rosângela Coelho.)*

The authors are with the Laboratory of Acoustic Signal Processing, Military Institute of Engineering (IME), Rio de Janeiro, RJ 22290-270, Brazil (e-mail: coelho@ime.eb.br).

Digital Object Identifier 10.1109/TASLP.2021.3093392

noises. Furthermore, the signal reconstruction does not require any knowledge regarding the phase of the target speech signal.

Several experiments are conducted to examine the effectiveness of the proposed solution. For this purpose, speech utterances are corrupted by four real acoustic noises with different impulsiveness degrees. Speech signals are corrupted with five values of signal-to-noise ratio (SNR): -10 dB, -5 dB, 0 dB, 5 dB, and 10 dB. Four speech enhancement techniques are adopted as baseline: the spectral Wiener filtering with unbiased minimum mean-square error estimator (UMMSE) [2], and the time-domain EMD-based filtering (EMDF) [13], the EMD-Hurst-based (EMDH) [9] approach, and the non-stationary noise estimation for speech enhancement (NNESE) [3]. The proposed HHT-α is evaluated considering seven objective measures that present high correlation with subjective listening tests. The perceptual evaluation of speech quality (PESQ) [18], the frequency-weighted segmental SNR (fwSNRseg) [19], the log-likelihood ratio (LLR), and the weighted spectral slope (WSS) are used to evaluate the enhanced signals in terms of speech quality. Regarding speech intelligibility, the short-time objective intelligibility measure (STOI) [20], the extended speech intelligibility index (ESII) [21], and the short-time variant of the approximated SII (ASII_{ST}) [22] are adopted to compare the proposed and baseline methods. Experiments demonstrate that the HHT-α method achieves the best speech quality results, especially for highly impulsive noises. HHT-α also shows interesting intelligibility scores when compared to the competitive techniques.

The main contributions of this study are:

- the introduction of the HHT-α speech enhancement solution to mitigate the effects of impulsive acoustic noises in the time domain;
- the definition of the impulsiveness index α as the criterion for impulsive noise detection;
- the adoption of the EEMD to avoid the use of the original noisy phase for the enhanced signal reconstruction;
- the description of a strategy for impulsive noise detection to yield quality and intelligibility improvement.

The remaining of this paper is organized as follows. Section II introduces the novel HHT-α speech enhancement method. Evaluation experiments to compare the proposed and baseline speech enhancement methods are presented in Section III, which also includes details of the noise database, and brief descriptions of the speech quality and intelligibility measures. Experiments results are presented in Section IV. Finally, Section V concludes this work.

II. HHT-α: Speech Enhancement Scheme

The HHT-α speech enhancement includes three main steps: noisy signal decomposition, estimation and selection of noise components, and speech signal reconstruction. Fig. 1 illustrates the block diagram of the proposed method.

A. Noisy Signal Decomposition

HHT [12] is a nonlinear adaptive approach that locally analyzes a signal $x(t)$ to define a local high-frequency part, also called detail $d(t)$, and a local trend $a(t)$, such that $x(t) = d(t) + a(t)$. An oscillatory IMF is derived from the detail function $d(t)$. The high- versus low-frequency separation procedure is iteratively repeated over the residual $a(t)$, leading to a new detail and a new residual. Thus, the decomposition leads to a series of IMFs and a residual, such that

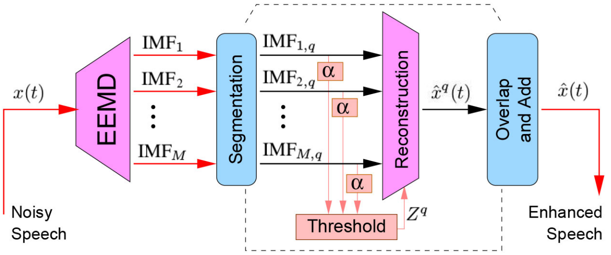

Fig. 1. Block diagram of the HHT-α speech enhancement method.

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

where $\text{IMF}_m(t)$ is the m-th mode of $x(t)$ and $r(t)$ is the residual. As opposed to other kinds of signal decomposition, a set of basis functions is not demanded for the HHT. In fact, HHT results in fully data-driven decomposition modes and does not require the stationarity of the target signal.

The EEMD was introduced in [16] to overcome the mode mixing problem that generally occurs in the original EMD. The key idea is to average IMFs obtained after corrupting the original signal using several realizations of white Gaussian noise (WGN). Thus, EEMD algorithm can be described as:

1) Generate $x^n(t) = x(t) + w^n(t)$, where $w^n(t)$, $n = 1, \ldots, N$, are different realizations of WGN;
2) Apply EMD to decompose $x^n(t)$, $n = 1, \ldots, N$, into a series of components $\text{IMF}^n_m(t)$, $m = 1, \ldots, M$;
3) Assign the m-th mode of $x(t)$ as

$$\text{IMF}_m(t) = \frac{1}{N} \sum_{n=1}^{N} \text{IMF}^n_m(t)\,; \qquad (1)$$

4) Finally, $x(t) = \sum_{m=1}^{M} \text{IMF}_m(t) + r(t)$, where $r(t)$ is the residual.

It is worth to mention that despite the lack of mathematical formulation, HHT is a very powerful tool for analyzing non-stationary real signals and has been successfully applied in several research areas [15], [23].

B. Estimation and Selection of Noise Components

In the literature, impulsive signals and noises are generally defined by a sequence of random samples with symmetric heavy-tail distribution, i.e., $P[X > x] \sim C|x|^{-\alpha}$, where C is a positive constant and $0 < \alpha \leq 2$ is the impulsiveness index. The α exponent is also related to α-stable distribution and may be described as the characteristic exponent [17].

In [24] authors showed that for α-stable noises the EMD behaves like a quasi-dyadic filterbank for $\alpha \in]1.2, 2.0]$. Speech signals investigated in this work are impulsive and present heavy-tails with α values in the range $[0.9, 1.2]$. On the other

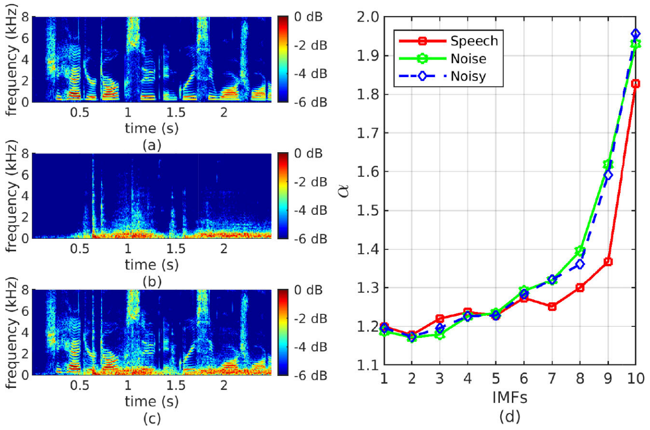

Fig. 2. Spectrograms of (a) clean speech (b) Sliding Door Closing noise ($\alpha = 1.21$), and (c) noisy speech (SNR=0 dB). (d) The average values of α estimated from the IMFs.

hand, acoustic noises commonly encountered in real urban scenarios have values in the range [1.2,2.0] [24]. Thus, in this paper the EMD is applied to highlight the noise impulsiveness of the corrupted speech signal. The estimator proposed by McCulloch in [25], [26] is here adopted for the α index estimation.

Fig. 2(a)-(c) show spectrograms of a clean speech signal collected from the TIMIT database [27], an impulsive Sliding Door Closing noise with $\alpha = 1.21$, and also the corrupted signal with SNR $= 0$ dB. Note from Fig. 2(b) that the noise energy is mostly concentrated at low frequencies and the spectrogram has sharp wide band components around 0.6, 1.0 and 1.5 seconds. Fig. 2(d) presents average values of the impulsiveness index α estimated from IMFs of clean speech, impulsive noise, and noisy speech signals. It can be seen that as the mode index increases, the α values of all signals approach 2. For the highest IMF indexes, e.g., $7-10$, the acoustic noise and the noisy speech signal have similar α values. These values are greater than those obtained from the clean speech signal. This indicates that these IMFs are more noise-like, which is in accordance with previous works (for example, refer to [9]). It is also interesting to note that for IMFs with indices $3-5$, the α values of the noisy signal vary between those estimated from the noise and from the clean speech signal. It means that the impulsiveness index is an appropriate identification criterion to select the IMFs with more speech-like characteristics and reject the noise-like components.

The selection of noise components is performed as follows. After the decomposition of the target noisy signal with the EEMD algorithm, each IMF is segmented into a set of overlapping short-time frames

$$\mathrm{IMF}_{m,q}(t) = \begin{cases} \mathrm{IMF}_m(t+qS_d)\ , & t \in [0, T_d], \\ 0 & , \text{elsewhere}, \end{cases} \qquad (2)$$

where $q \in \{0, \ldots, Q-1\}$ is the frame index, T_d is the fixed time-duration of the frames in samples, and S_d is the step size in samples.

In this proposal, the selection of noisy components is based on α parameters of each windowed IMF. For each frame q, the impulsiveness index is estimated from the decomposition modes $\mathrm{IMF}_{m,q}(t)$ leading to a set of values $\alpha_1^q, \ldots, \alpha_M^q$. The next step is to determine the index Z^q of the last IMF whose impulsiveness index is below a given threshold, ρ_α, i.e., $\alpha_Z^q \leq \rho_\alpha$. IMFs whose α values exceed the threshold are considered as noise-like components.

C. Speech Signal Reconstruction

If $\hat{x}(t)$ represents the enhanced speech signal, then each frame, $\hat{x}^q(t)$, is reconstructed by

$$\hat{x}^q(t) = \sum_{m=1}^{Z^q} w(t)\, \mathrm{IMF}_{m,q}(t), \quad q = 0, \ldots, Q-1, \qquad (3)$$

where Z^q is the index of the last mode considered as speech and $w(t)$ is a window function used to avoid discontinuities in the reconstructed signal (for more details see [9]). Finally, $\hat{x}(t)$ is reconstructed by overlapping and adding all the frames $\hat{x}^q(t), q = 0, \ldots, Q-1$, as

$$\hat{x}(t) = \frac{1}{P} \sum_{q=1}^{Q} \hat{x}^q(t - qS_d)\,, \qquad (4)$$

where P is a normalization factor that depends on the window function $w(t)$, the frame length T_d, and the step size S_d. It is worth mentioning that, unlike frequency-domain methods, this reconstruction procedure does not require the phase information from the original noisy signal. Moreover, the proposed speech enhancement solution does not impose any constraint regarding the distribution or the stationarity of the noise and speech signals.

III. Speech Enhancement Evaluation Experiments

Extensive speech enhancement experiments are conducted to examine the HHT-α method in terms of speech quality and speech intelligibility. These experiments consider a subset of 20 speech segments of the TIMIT speech database [27]. Speech utterances have sampling rate of 16 kHz and time duration of 2.5 seconds. The time-domain EMDF, EMDH, and NNESE methods, and also the spectral UMMSE are adopted as baseline for the evaluation of the proposed solution. Experiments are conducted considering noisy speech signals with five SNR values: -10 dB, -5 dB, 0 dB, 5 dB, and 10 dB.

A. Noise Database

Four real impulsive acoustic noises are used to corrupt the speech utterances: Sliding Door Closing ($\alpha = 1.21$), Industrial Machine ($\alpha = 1.40$), and Horn ($\alpha = 1.59$) are selected from Freesound.org,[1] while Babble ($\alpha = 1.79$) is obtained from the RSG-10 [28] database. These noise files are also available at lasp.ime.eb.br. Fig. 3 presents the spectrogram and the index of non-stationarity (INS) [29] obtained from segments of the acoustic noises. The INS values (blue plots) are here shown to objectively examine the non-stationarity of impulsive noises. 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 ($T = 2.5$ seconds) of noise sample sequences. For each window length T_h,

[1] [Online]. Available: https://freesound.org

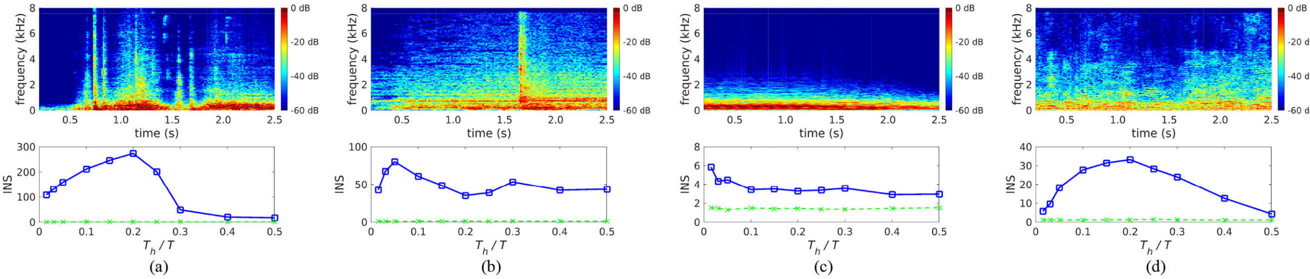

Fig. 3. Spectrograms (upper maps) and INS (blue lines in lower part) obtained for 2.5-seconds segments of the acoustic impulsive noises: (a) Sliding Door Closing, (b) Industrial Machine, (c) Horn, and (d) Babble. Green dashed lines in lower part indicate the value for the stationarity test threshold.

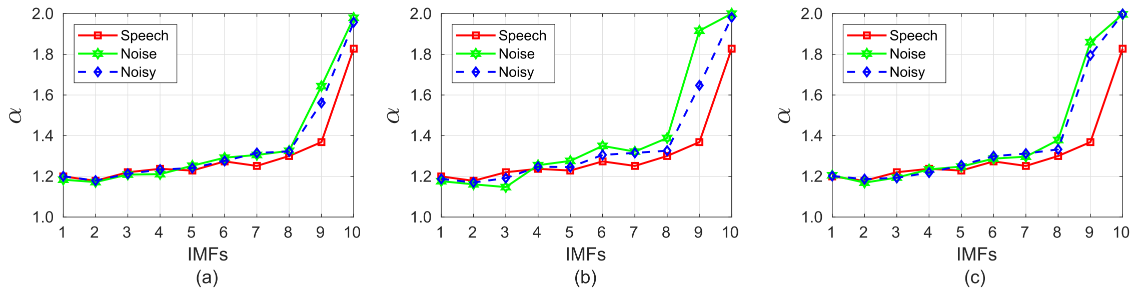

Fig. 4. Average values of α estimated from the IMFs considering impulsive noise sources: (a) Industrial Machine, (b) Horn, and (c) Babble.

a threshold is defined to guarantee the stationarity assumption with a confidence degree of 95%. Thus, if INS $\leq \gamma$ then the noise is considered as stationary. Otherwise, it is designated as non-stationary. The γ values (green plots) are also depicted in Fig. 3. Sliding Door Closing, Industrial Machine, and Babble noises are here classified as highly non-stationary since their INS achieves values greater than 200, 80, and 30, respectively. Horn noise exhibits INS results in the range $[3, 6]$ and thus, it is defined as moderately non-stationary.

As a complement of Fig. 2(d), Fig. 4 depicts the values of α estimated from different IMFs considering the other three impulsive noises: Industrial Machine, Horn, and Babble. The noisy signals are obtained considering SNR of 0 dB. Note that once again α values indicate that IMFs with high indices are mostly composed by noise. It reinforces the comments in Section II-B: the impulsiveness index α is an interesting criterion to detect the most corrupted IMFs.

B. Speech Quality Measures

Four objective measures are adopted to evaluate the proposed method in terms of speech quality: PESQ [18], fwSNRseg, LLR, and WSS. These measures present high correlation with subjective overall quality and signal distortion results [19].

1) PESQ: Due to the impulsive nature of the acoustic noises, in this work a modified version of PESQ is adopted for the evaluation of the enhanced signals. It means that the symetric disturbance ($d_{s,q}$) and the asymetric disturbance ($d_{a,q}$) are estimated from each frame q to achieve a frame PESQ score given by

$$\text{PESQ}_q = 4.5 - 0.1 d_{s,q} - 0.0309 d_{a,q}\,. \tag{5}$$

When corrupted by highly impulsive noises, speech signals may be severely disturbed at certain time instants while no perceptive corruption are found in other regions. In such scenario, the severely corrupted segments are the ones that most contribute to speech quality and intelligibility degradation. In order to prove that the HHT-α method is able to detect and compensate the acoustic noise from these regions, the scores computed with (5) are averaged considering only frames that are highly affected by the acoustic noise. The PESQ score is here computed from 50% of the frames based on an SNR criterion, namely from those with the lowest SNR values.

2) fwSNRseg: The frequency-weighted segmental SNR is defined as a weighted sum of SNR values computed in 25 frequency bands using Gaussian-shaped filters. In this work, the fwSNRseg values are computed considering the weighting function introduced in [19] to achieve high correlation with quality results from perceptual listening tests. Results described in [30] also demonstrated that the fwSegSNR is highly correlated to subjective speech intelligibility scores.

3) LLR: The log-likelihood ratio is computed based on the linear prediction coefficients (LPC) obtained from the clean and processed versions of the speech signal. As defined in [31], the LLR of each frame q is computed as

$$\text{LLR}(q) = \log\left(\frac{\boldsymbol{a}_{p,q}\,\mathbf{R}_c\,\boldsymbol{a}_{p,q}^T}{\boldsymbol{a}_{c,q}\,\mathbf{R}_c\,\boldsymbol{a}_{c,q}^T}\right), \tag{6}$$

where $\boldsymbol{a}_{c,q}$ and $\boldsymbol{a}_{p,q}$ are the LPC vectors from the original and processed frames, respectively, and $\mathbf{R}_c$ is the autocorrelation matrix of the clean speech signal. A single LLR value is then obtained by averaging the smallest 95% of the frame LLR values limited to the range $[0, 2]$.

4) WSS: The weighted spectral slope was proposed in [32] as a distance measure based on the idea that speech quality is highly affected by differences in vowels formant frequencies. The WSS is then computed as a weighted sum of the differences between the spectral slopes of the clean and enhanced versions of the speech signal. The weight of each frequency band is able to heavily penalize large differences and ignore small variations between clean and enhanced spectra.

C. Speech Intelligibility Measures

The proposed HHT-α is also evaluated in terms of speech intelligibility. STOI [20], ESII [21], and ASII$_{\text{ST}}$ [22] measures are adopted for speech intelligibility assessment. Similarly to the criterion adopted for the PESQ computation, the intelligibility scores also consider the 50% frames most affected by the impulsive acoustic noises.

1) STOI: The short-time objective intelligibility measure was proposed in [20] as a correlation-based method to compare the spectrum of the clean and the enhanced speech signals in the frequency domain. The correlation between temporal envelopes of the clean and noisy speech signals is defined as the intermediate intelligibility measure $\text{STOI}_{(j,q)}$ of each frequency band j and each time frame q. The STOI is finally given by

$$\text{STOI} = \frac{1}{15Q} \sum_{q=1}^{Q} \sum_{j=1}^{15} \text{STOI}_{(j,q)}, \tag{7}$$

where Q is the number of speech frames.

2) ESII: The extended speech intelligibility index was proposed in [21] as a short-time adaptation of the SII defined in ANSI S3.5-1997 [33]. For the ESII computation, the SNR $\xi(j,q)$ values computed at each critical frequency band j and time frame q are first normalized and clipped to the range $[0,1]$ by

$$d(j,q) = \frac{\max(\min(10\log_{10}\xi(j,q), 15), -15)}{30} + \frac{1}{2}\,. \tag{8}$$

The ESII is then computed as a weighted average of all values given in (8):

$$\text{ESII} = \frac{1}{Q} \sum_{q=1}^{Q} \sum_{j=1}^{J} \gamma_j \, d(j,q)\,, \tag{9}$$

where J is the total number of critical bands, and γ_i are the critical-band-importance weights.

3) ASII$_{ST}$: In the short-time variant of the approximated SII [22], the function $d(j,q)$ adopted in (8) to normalize the SNR $\xi(j,q)$ is replaced by

$$d(j,q) = \frac{\xi(j,q)}{\xi(j,q)+1}\,. \tag{10}$$

The ASII$_{\text{ST}}$ score is here computed using the same weights γ_i adopted in the ESII.

D. Definition of the Noise Selection Threshold

In the proposed HHT-α method, the decision threshold ρ_α is crucial to determine the components to be removed from each corrupted speech frame. In this work, an adaptive threshold is introduced for the selection criterion, such that

$$\rho_\alpha = \max(\xi\alpha_u^q, \alpha_{\min})\,, \tag{11}$$

where α_u^q is the estimate of α for the corrupted speech windowed signal. The parameter ξ is adopted to adjust the amount of noise components to be removed, while $\alpha_{\min}$ is set to 1.1 to avoid excessive component removal in speech dominant segments of the signal.

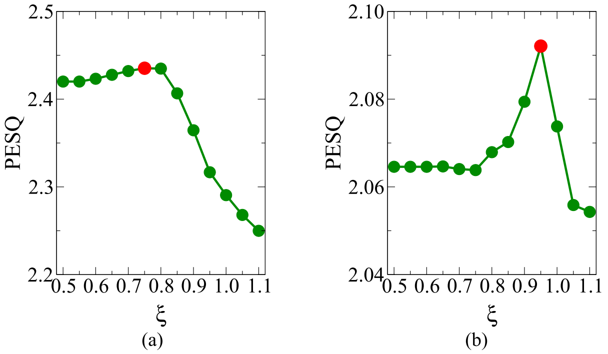

Fig. 5. Average PESQ computed from speech signals enhanced by the HHT-α method considering the selection threshold with different values of ξ. Speech signals are corrupted by (a) Sliding Door Closing and (b) Industrial Machine noises considering SNR of 0 dB.

Preliminary experiments are conducted in order to empirically determine the optimal value of ξ to be adopted in (11). For this purpose, two noise sources are used to corrupt the noisy speech signals considering SNR of 0 dB. Sliding Door Closing and Industrial Machine are selected since they are highly impulsive noises, i.e., present the lowest α values. For the HHT-α implementation, the EEMD is applied considering 50 different realizations of WGN with SNR of 30 dB to obtain 10 IMFs. The selection of noise components considers $T_d = 10240$ samples per frame and step size of $S_d = 128$ samples. The HHT-α method is applied to these signals with values of ξ varying in the range $[0.5, 1.1]$. This range is defined according to the importance of α_u^q and $\alpha_{\min}$. When the value of ξ approaches 0, α_u^q is not taken into account since $\rho_\alpha = \alpha_{\min}$. On the other hand, values of ξ greater than 1 would lead the threshold ρ to achieve too high values, which may result in excessive component removal in speech dominant segments of the signal. The average PESQ results are depicted in Fig. 5. For these acoustic noises, the highest PESQ values are attained with $\xi \in [0.75, 0.95]$. Thus, the value $\xi = 0.85$ is adopted in all the following experiments to evaluate the proposed HHT-α method in terms of speech quality and intelligibility. The ξ value was evaluated considering a rich manifold scenario, i.e., 20 different noisy conditions (4 noises, 5 SNR values). The main idea of using a single value is to guarantee a diversity of unbiased testing conditions.

IV. Results and Discussion

This Section presents the speech quality and intelligibility results obtained with the proposed HHT-α and baseline speech enhancement techniques. An example of the HHT-α method applied to a noisy signal is illustrated in Fig. 6, which considers

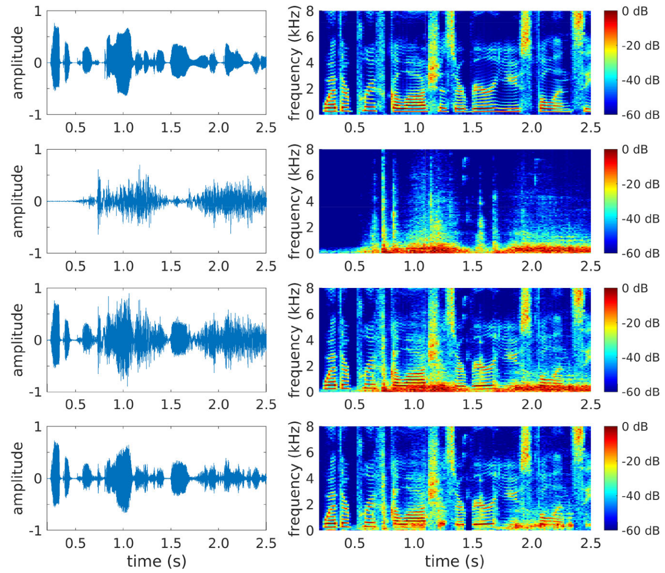

Fig. 6. Time domain amplitudes and spectrograms of an example speech signal: (a) clean speech signal, (b) Sliding Door Closing noise, (c) noisy signal considering SNR of 0 dB, and (d) the enhanced signal obtained with HHT-α.

the Sliding Door Closing noise with SNR of 0 dB. Fig. 6 depicts time-domain amplitudes and spectrograms of the clean signal, the acoustic noise, the noisy signal, and the enhanced signal. It is possible to note that HHT-α was able to detect most of the noise content from the noisy signal.

A. Speech Quality Evaluation

Table I shows the PESQ results obtained with the proposed and competing speech enhancement techniques. Note that HHT-α outperforms the competing time and spectral domain approaches for most of the noisy scenarios. The proposed solution achieves the highest PESQ values in 19 of 20 noisy conditions. It can be seen that the proposed HHT-α attains interesting results for the most impulsive noise. For instance, the PESQ scores are about 0.2 higher than those achieved by the competing solutions for SNR ≥ 0 dB. On average, the overall PESQ obtained with the HHT-α is 2.32, which is 0.09, 0.12, 0.13, and 0.14 higher than UMMSE, EMDH, EMDF, and NNESE, respectively. Furthermore, HHT-α also achieves the best scores for the Industrial Machine noise. This result is particularly important due to the small fluctuations depicted in Fig. 5(b), which means that PESQ is insensitive to the value of ξ for this noise.

Fig. 7 exhibits the average fwSNRseg improvement obtained by the proposed and baseline methods for the four noises considering SNR of -10 dB, 0 dB, and 10 dB. It is interesting to mention that HHT-α achieves the best results for three noise sources considering SNR values of -10 dB and 0 dB. When compared to the time-domain EMDF, EMDH, and NNESE approaches, the proposed solution achieves the highest fwSNRseg values for almost all noise conditions. For the highly impulsive Sliding Door Closing noise, HHT-α also outperforms the spectral UMMSE. For the other noise sources, HHT-α results are superior than UMMSE for the lowest SNR values, while UMMSE attains the highest improvement for SNR of 10 dB. Note that for the Horn noise, the HHT-α and UMMSE techniques achieve highly significant fwSNRseg results. This means that HHT-α and UMMSE are better able to satisfactorily detect the low frequency content of noise.

TABLE I
PESQ Results With the Proposed and Baseline Methods

Noise	SNR	UMMSE	EMDF	EMDH	NNESE	HHT-α
Sliding Door Closing $\alpha = 1.21$	-10 dB	1.69	1.70	1.69	1.69	**1.73**
	-5 dB	1.92	1.93	1.92	1.92	**2.06**
	0 dB	2.23	2.23	2.23	2.22	**2.43**
	5 dB	2.57	2.56	2.57	2.55	**2.78**
	10 dB	2.93	2.92	2.93	2.91	**3.12**
	Average	2.27	2.27	2.27	2.26	**2.42**
Industrial Machine $\alpha = 1.40$	-10 dB	1.57	1.57	1.58	1.56	**1.70**
	-5 dB	1.79	1.81	1.81	1.80	**1.89**
	0 dB	2.03	2.02	2.02	2.02	**2.08**
	5 dB	2.33	2.30	2.30	2.30	**2.34**
	10 dB	**2.72**	2.66	2.66	2.66	**2.72**
	Average	2.09	2.07	2.07	2.07	**2.15**
Horn $\alpha = 1.59$	-10 dB	2.00	1.97	1.96	1.95	**2.03**
	-5 dB	2.22	2.12	2.12	2.11	**2.31**
	0 dB	2.49	2.36	2.37	2.35	**2.62**
	5 dB	2.82	2.64	2.65	2.62	**2.93**
	10 dB	3.19	2.96	2.97	2.94	**3.24**
	Average	2.54	2.41	2.41	2.39	**2.63**
Babble $\alpha = 1.79$	-10 dB	1.56	1.65	1.66	1.64	**1.71**
	-5 dB	1.69	1.70	1.70	1.69	**1.80**
	0 dB	1.96	1.94	1.94	1.94	**2.02**
	5 dB	2.28	2.23	2.24	2.24	**2.31**
	10 dB	**2.66**	2.58	2.59	2.58	2.64
	Average	2.03	2.02	2.03	2.02	**2.10**
Overall		2.23	2.19	2.20	2.18	**2.32**

Fig. 8 depicts the average LLR results obtained for each acoustic noise. Once again, the proposed and baseline methods are compared considering SNR of -10 dB, 0 dB, and 10 dB. Similar to the findings with PESQ, the HHT-α achieves the highest LLR for the Industrial Machine and Babble noises. It also outperforms the competing methods for the Sliding Door Closing noise with SNR of 10 dB. The NNESE attains the best LLR results for the remaining noise scenarios. From the total of 12 noise conditions, the proposed solution outperforms the spectral UMMSE for 10 situations. For the Sliding Door Closing noise with SNR of -10 dB and 0 dB, HHT-α and UMMSE attain the same average LLR values. The overall LLR obtained with HHT-α is 0.73, while results achieved with NNESE is 0.75 and the other approaches attain 0.66 or less.

Fig. 9 presents the average WSS values achieved by the proposed and baseline methods. As a distance measure, the best quality results correspond to the lowest WSS values. Note that the proposed HHT-α attains the lowest WSS for almost all noise scenarios with SNR ≤ 0 dB. For SNR of 10 dB, HHT-α achieves the best results for Sliding Door Closing and Horn

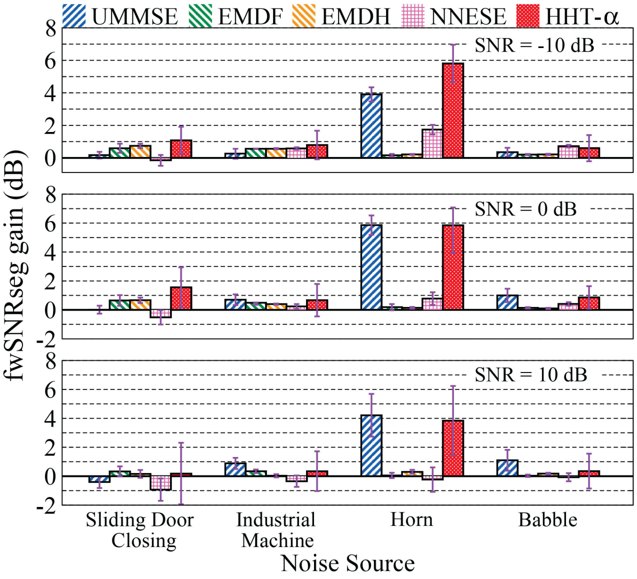

Fig. 7. Average fwSNRseg gain obtained for different noise sources considering SNR of −10 dB (top), 0 dB (middle), and 10 dB (bottom).

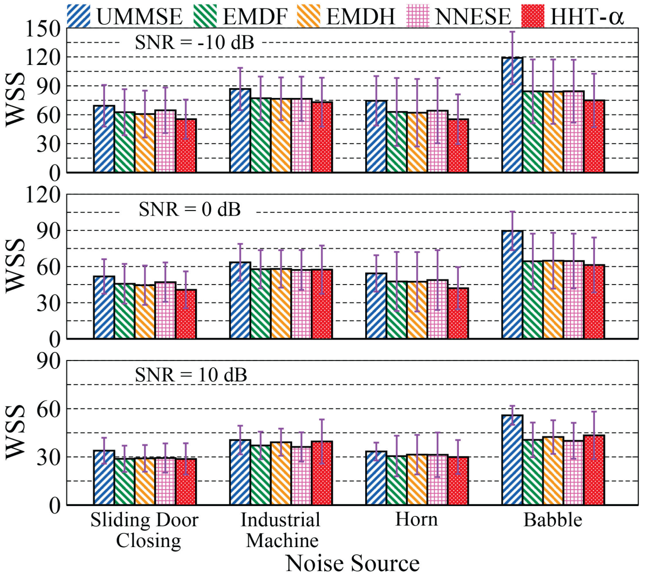

Fig. 9. Average WSS results obtained for different noise sources considering SNR of −10 dB (top), 0 dB (middle), and 10 dB (bottom).

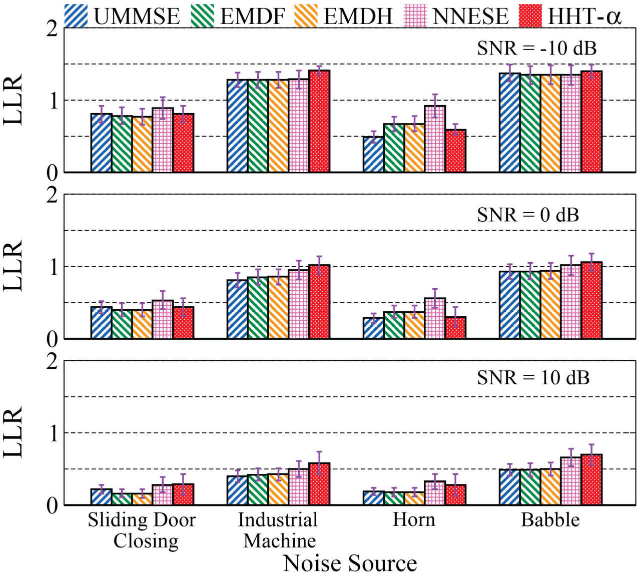

Fig. 8. Average LLR results obtained for different noise sources considering SNR of −10 dB (top), 0 dB (middle), and 10 dB (bottom).

noises. Moreover, the proposed solution outperforms the spectral UMMSE for all noise conditions.

B. Speech Intelligibility Evaluation

Table II presents intelligibility scores obtained with STOI. Note that HHT-α achieves the best results in 18 from a total of 20 noise conditions. Once again, the performance of the proposed method is particularly interesting for the most impulsive noises, i.e., Sliding Door Closing and Industrial Machine. For these noise sources, the STOI scores attained by HHT-α are considerably higher than all the competing solutions for all SNR values. The spectral UMMSE attains the highest scores in only two situations: speech signals corrupted by Horn and Babble noises with SNR of 10 dB. However, even in these scenarios the proposed solution outperforms the other time-domain methods. On average, the proposed approach attains an intelligibility score of 0.71, which is 0.08 greater than UMMSE. The time-domain NNESE, EMDH, and EMDF solutions achieve average STOI scores of 0.62, 0.61, and 0.59, respectively.

The ASII$_{\mathrm{ST}}$ and ESII measures are also considered for the evaluation of the HHT-α in terms of speech intelligibility. As a reference for the objective intelligibility scores, the average ESII and ASII$_{\mathrm{ST}}$ values obtained from the noisy (unprocessed) speech signals are shown in Tables III and IV, respectively. The values 0.45 and 0.75 are considered as thresholds for poor and good intelligibility, respectively [33], [34]. It means that variations in the intelligibility measures outside the range $[0.45, 0.75]$ do not lead to practical intelligibility changes.

Fig. 10 illustrates the ESII improvement (ΔESII) obtained with the proposed and baseline methods for the four acoustic noises. Note that HHT-α achieves the highest intelligibility gain for the noises with the lowest α values: Sliding Door Closing, Industrial Machine, and Horn. The proposed method is able to improve the intelligibility results even in those situations where the competing methods achieve negative gain: Sliding Door Closing with SNR ≥ 5 dB and Horn with SNR ≥ -5 dB. For the Babble noise, the proposed solution still outperforms the time-domain EMDF, EMDH, and NNESE solutions.

The average ASII$_{\mathrm{ST}}$ gain (ΔASII$_{\mathrm{ST}}$) obtained with the proposed and baseline techniques are depicted in Fig. 11. Once again, the superior performance of the HHT-α solution is more prominent for the Sliding Door Closing, Industrial Machine, and Horn noises. Moreover, HHT-α improves the ASII$_{\mathrm{ST}}$ results even in those scenarios where the spectral UMMSE leads to negative gain (refer to Fig. 11(a),(c)). From Tables III and IV and from the improvement results respectively illustrated in Figs. 10 and 11, it is possible to verify noticeable intelligibility achievement. For

TABLE II
STOI SPEECH INTELLIGIBILITY SCORES

Noise	SNR	UMMSE	EMDF	EMDH	NNESE	HHT-α
Sliding Door Closing $\alpha = 1.21$	-10 dB	0.45	0.46	0.47	0.47	**0.55**
	-5 dB	0.57	0.57	0.57	0.57	**0.70**
	0 dB	0.66	0.67	0.67	0.66	**0.81**
	5 dB	0.75	0.75	0.75	0.77	**0.87**
	10 dB	0.82	0.82	0.82	0.85	**0.89**
	Average	0.65	0.65	0.66	0.66	**0.76**
Industrial Machine $\alpha = 1.40$	-10 dB	0.31	0.32	0.33	0.33	**0.34**
	-5 dB	0.40	0.40	0.42	0.42	**0.46**
	0 dB	0.52	0.51	0.53	0.53	**0.66**
	5 dB	0.65	0.64	0.65	0.69	**0.81**
	10 dB	0.78	0.76	0.77	0.84	**0.87**
	Average	0.53	0.53	0.54	0.56	**0.63**
Horn $\alpha = 1.59$	-10 dB	0.61	0.53	0.56	0.55	**0.69**
	-5 dB	0.69	0.62	0.63	0.62	**0.77**
	0 dB	0.77	0.71	0.72	0.72	**0.80**
	5 dB	0.85	0.78	0.78	0.79	**0.86**
	10 dB	**0.91**	0.84	0.84	0.86	0.87
	Average	0.77	0.70	0.71	0.71	**0.80**
Babble $\alpha = 1.79$	-10 dB	0.31	0.31	0.32	0.32	**0.38**
	-5 dB	0.38	0.36	0.38	0.38	**0.50**
	0 dB	0.51	0.46	0.49	0.49	**0.68**
	5 dB	0.71	0.61	0.63	0.73	**0.81**
	10 dB	**0.88**	0.76	0.78	0.86	0.87
	Average	0.56	0.50	0.52	0.56	**0.65**
Overall		0.63	0.59	0.61	0.62	**0.71**

TABLE III
ESII RESULTS FOR UNPROCESSED SPEECH SIGNALS

Noise	SNR				
	-10 dB	-5 dB	0 dB	5 dB	10 dB
Sliding Door Closing	0.30	0.37	0.46	0.55	0.64
Industrial Machine	0.14	0.19	0.26	0.35	0.45
Horn	0.45	0.52	0.59	0.66	0.73
Babble	0.12	0.17	0.24	0.32	0.42
Helicopter	0.19	0.26	0.34	0.43	0.53

TABLE IV
ASII$_{ST}$ RESULTS FOR UNPROCESSED SPEECH SIGNALS

Noise	SNR				
	-10 dB	-5 dB	0 dB	5 dB	10 dB
Sliding Door Closing	0.26	0.34	0.45	0.56	0.67
Industrial Machine	0.09	0.14	0.22	0.32	0.44
Horn	0.45	0.52	0.60	0.68	0.75
Babble	0.07	0.11	0.18	0.28	0.40

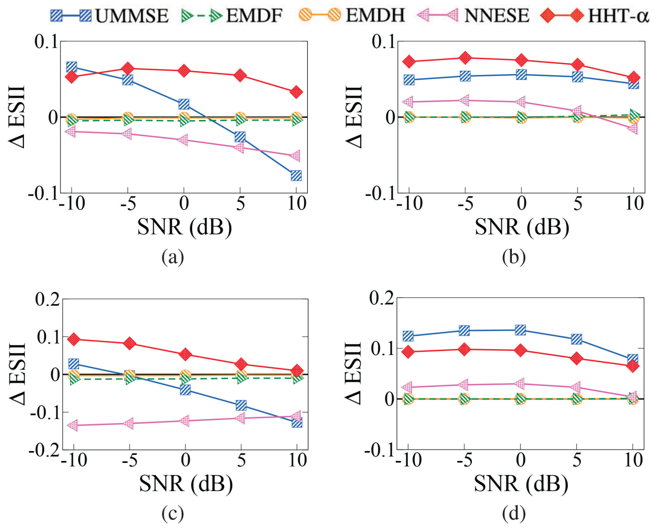

Fig. 10. ESII improvement (ΔESII) obtained for (a) Sliding Door Closing, (b) Industrial Machine, (c) Horn, and (d) Babble noises.

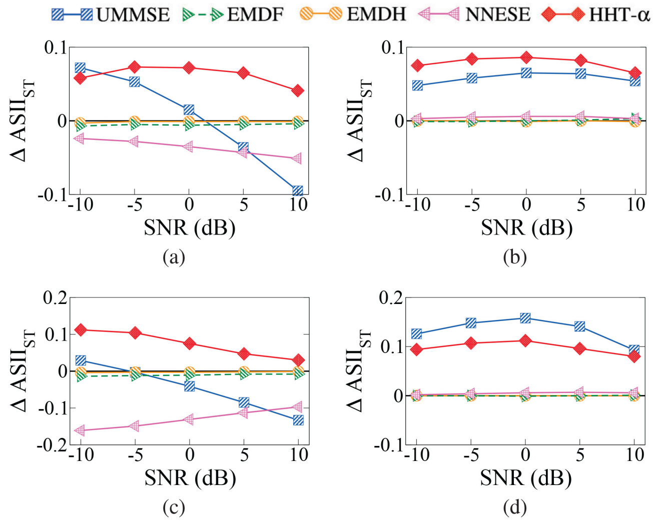

Fig. 11. ASII $_{ST}$ improvement (Δ ASII$_{ST}$) obtained for (a) Sliding Door Closing, (b) Industrial Machine, (c) Horn, and (d) Babble noises.

example, for the Sliding Door Closing, SNR $=0$ dB, the HHT-α increases the intelligibility from 0.45 to 0.52 for the ASII$_{ST}$.

In summary, the proposed HHT-α leads to the best speech quality results considering the four objective measures: PESQ, fwSNRseg, LLR, and WSS. Some exceptions can be found with Horn and Babble noises with SNR ≥ 0 dB, where the UMMSE achieved slightly superior fwSNRseg results. Another particular situation concerns the LLR measure, where the NNESE attained the best results for the Horn noise. In terms of speech intelligibility, the HHT-α reaches the highest STOI, ESII, and ASII$_{ST}$ intelligibility scores. The exception is the Babble noise, for which the UMMSE obtains the highest ESII and ASII$_{ST}$ values. For the highly impulsive Sliding Door Closing noise, HHT-α achieves the best results in terms of both objective and intelligibility measures (refer to Table I and Figs. 7 and 8). The intelligibility improvement scores with the proposed HHT-α also highlight the source separation aspect as function of its impulsive noise detection criterion.

V. Conclusion

This paper introduced the HHT-α speech enhancement technique based on the Hilbert-Huang Transform. The EEMD algorithm was applied to decompose the noisy speech signal in the time domain. The estimation and selection of noise components was performed frame-by-frame based on the impulsiveness index of the decomposition modes. The enhanced version of the speech signal was finally reconstructed using the IMFs that are mainly composed of speech. Several experiments were conducted to evaluate the proposed method with the UMMSE, EMDF, EMDH, and NNESE competitive speech enhancement solutions. Four non-stationary acoustic noises with different the impulsiveness indices were adopted for this purpose. Objective quality results demonstrated that the proposed HHT-α leads to superior speech quality when compared to the baseline approaches. Particularly for the most impulsive noise, the proposed solution outperformed the four competing approaches in terms of PESQ, fwSNRseg, and WSS objective quality measures. For the objective speech intelligibility STOI measure, HHT-α achieved superior scores for almost all noisy conditions. The ESII and ASII_{ST} objective measures were also applied to reinforce the significant improvement in terms of speech intelligibility. Once again, the HHT-α attained the best quality and intelligibility scores for the highly impulsive Sliding Door Closing. Future research includes the investigation of the impulsive noise effect on human auditory system [1] and its relationship with intelligibility impairment.

References

[1] R. Davis and O. Clavier, "Impulsive noise: A brief review," *Hear. Res.*, vol. 349, pp. 34–36, 2017.

[2] 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.

[3] R. Tavares and R. Coelho, "Speech enhancement with nonstationary acoustic noise detection in time domain," *IEEE Signal Process. Lett.*, vol. 23, no. 1, pp. 6–10, Jan. 2016.

[4] S. M. Nørholm, J. R. Jensen, and M. G. Christensen, "Enhancement and noise statistics estimation for non-stationary voiced speech," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 24, no. 4, pp. 645–658, Apr. 2016.

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

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

[7] M. Ruhland, J. Bitzer, M. Brandt, and S. Goetze, "Reduction of Gaussian, supergaussian, and impulsive noise by interpolation of the binary mask residual," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 10, pp. 1680–1691, Oct. 2015.

[8] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, "Phase processing for single-channel speech enhancement," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 55–66, Mar. 2015.

[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, no. 5, 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 Press, 2015.

[11] J. R. Jensen, J. Benesty, M. G. Christensen, and S. H. Jensen, "Enhancement of single-channel periodic signals in the time-domain," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 7, pp. 1948–1963, Sep. 2012.

[12] N. Huang *et al.*, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," in *Proc. Roy. Soc. London Ser. A: Math., Phys., Eng. Sci.*, vol. 454, no. 1971, Mar. 1998, pp. 903–995.

[13] 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.

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

[15] A. Stallone, A. Cicone, and M. Materassi, "New insights and best practices for the successful use of empirical mode decomposition, iterative filtering and derived algorithms," *Nat. Sci. Rep.*, vol. 10, no. 15161, pp. 1–15, 2020.

[16] M. E. Torres, M. A. Colominas, G. Schlotthauer, and P. Flandrin, "A complete ensemble empirical mode decomposition with adaptive noise," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2011, pp. 4144–4147.

[17] C. Nikias and M. Shao, *Signal Processing With Alpha-Stable Distributions and Applications*. Hoboken, NJ, USA: Wiley, 1995.

[18] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, May 2001, pp. 749–752.

[19] 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.

[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] K. Rhebergen and N. Versfeld, "A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," *J. Acoust. Soc. Amer.*, vol. 117, no. 4, pp. 2181–2192, 2005.

[22] R. Hendriks, J. Crespo, J. Jensen, and C. Taal, "Optimal near-end speech intelligibility improvement incorporating additive noise and late reverberation under an approximation of the short-time SII," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 23, no. 5, pp. 851–862, May 2015.

[23] N. Huang, "Introduction to the hilbert-huang transform and its related mathematical problems," in *Hilbert-Huang Transform and its Applications*, N. Huang and S. Shen, Eds. Singapore: World Sci. Publishing, 2014.

[24] A. Komaty, A. Boudraa, J. Nolan, and D. Dare, "On the behavior of EMD and MEMD in presence of symmetric alpha-stable noise," *IEEE Signal Process. Lett.*, vol. 22, no. 7, pp. 818–822, Jul. 2015.

[25] J. H. McCulloch, "Simple consistent estimators of stable distribution parameters," *Commun. Statist. - Simul. Comput.*, vol. 15, no. 5, pp. 1109–1136, 1986.

[26] J. H. McCulloch, "Maximum likelihood estimation of symmetric stable parameters," Ohio State Univ., Dept. Econ., Tech. Rep., 1998.

[27] J. Garofolo *et al.*, "TIMIT acoustic-phonetic continuous speech corpus," in *Linguist. Data Consortium*, Philadelphia, PA, USA, 1993.

[28] H. Steeneken and F. Geurtsen, "Description of the RSG-10 noise database," Tech. Rep., TNO Institute for Perception, The Netherlands, 1988.

[29] 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.

[30] 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, 2009.

[31] S. Quackenbush, T. Barnwell, and M. Clements, *Objective Measures of Speech Quality*. Englewood Cliffs, NJ, USA: Prentice-Hall, Inc., 1988.

[32] D. Klatt, "Prediction of perceived phonetic distance from critical-band spectra: A first step," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 7, pp. 1278–1281, May 1982.

[33] *Methods for calculation of the speech intelligibility index*, ANSI S3.5–1997, American National Standard Institute, 1997.

[34] B. Sauert and P. Vary, "Near end listening enhancement: Speech intelligibility improvement in noisy environments," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. Proc.*, vol. 1, 2006, pp. 493–496.

César Medina received the B.Sc. degree in electrical engineering from the Escuela Politécnica del Ejército, Quito, Ecuador, in 2000, the M.Sc. degree from the Military Institute of Engineering (IME), Rio de Janeiro, Brazil in 2003, and the Ph.D. degree from the Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Rio de Janeiro, Brazil, in 2009. He is currently a Senior Leader of the Artificial Intelligence and Data Science Group, Radix in Rio de Janeiro, Brazil. From 2009 to 2014 and from 2015 to 2016, he was a Postdoctoral Researcher with PUC-Rio, and from 2016 to 2018, with IME. His research interests include signal processing for communications, adaptive systems, array processing, acoustic signal processing, speech and music processing, and speech enhancement.

Rosângela Coelho (Senior Member, IEEE) received the M.Sc. degree in electrical engineering from the Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Rio de Janeiro, Brazil, and the Ph.D. degree in electrical engineering from the École Nationale Supérieure des Télécommunications (Télécom Paris), Paris, France. In 2002, she joined the Military Institute of Engineering (IME), Rio de Janeiro, Brazil, and she is currently a Professor of the Electrical Engineering Section. She also heads the Laboratory of Acoustic Signal Processing, IME. From 1996 to 2001, she was a Postdoctoral Researcher with PUC-Rio, Télecom Paris and IME. Her current research interests include acoustic signal processing, speech enhancement, time-frequency acoustic masks, noise and reverberation detection, localization in urban acoustic scenes, and wireless acoustic sensor networks. She is currently Senior Member of the IEEE Signal Processing Society, and an Affiliate Member of the Technical Committee Audio and Acoustic Signal Processing. Since 2018, she has been an Associate Editor for the IEEE SIGNAL PROCESSING LETTERS. From 2000 to 2007, she was also an Editorial Board Member of the IEEE COMMUNICATIONS SURVEYS AND TUTORIALS. Since 2018, she has been an Associate Editor for the IEEE SIGNAL PROCESSING LETTERS.

Leonardo Zão (Member, IEEE) received the B.Sc. degree in electrical engineering, and the M.Sc. and Ph.D. degrees from the Military Institute of Engineering, Rio de Janeiro, Brazil, in 2005, 2010, and 2013, respectively. He is currently an Assistant Professor with the Military Institute of Engineering. Since 2014, he has been a Research Assistant with the Laboratory of Acoustic Signal Processing. His current research interests include acoustic signal processing, speaker recognition, speech enhancement, acoustic mask for classification, and adaptive learning.