ESPRIT-Hilbert based audio tampering detection with SVM classifier for forensic analysis via electrical network frequency

Paulo Max G. I. Reis, João Paulo C. L. da Costa, *Senior Member, IEEE*, Ricardo K. Miranda, *Student Member, IEEE* and Giovanni Del Galdo, *Member, IEEE*

***Abstract*—Audio authentication is a critical task in multimedia forensics demanding robust methods to detect and identify tampered audio recordings. In this article, a new technique to detect adulterations in audio recordings is proposed by exploiting abnormal variations in the Electrical Network Frequency (ENF) signal eventually embedded in a questioned audio recording. These abnormal variations are caused by abrupt phase discontinuities due to insertions and suppressions of audio snippets during the tampering task. First we propose an ESPRIT-Hilbert ENF estimator in conjunction with an outlier detector based on the sample kurtosis of the estimated ENF. Next we use the computed kurtosis as input for a Support Vector Machine (SVM) classifier to indicate the presence of tampering. The proposed scheme, herein designated as SPHINS, significantly outperforms the state-of-the-art high precision phase detection approach in the tests made. We validate our results using the Carioca 1 corpus with 100 unedited authorized audio recordings of phone calls.**

***Index Terms*—Acoustical signal processing, tampering detection, ESPRIT, Audio authenticity, electric network frequency (ENF)**

I. Introduction

The exponential increase in access to many sophisticated communication technologies makes multimedia evidences to have a key role in criminal prosecution. Particularly in the last ten years, there has been a significant growth of digital audio recordings being presented as evidence in crime related investigations and trials. Records from many cases involving eavesdropping devices are often the main if not the only proof of the materiality and authorship of a specific crime.

The large scale use of digital audio recordings became easier nowadays due to the low cost of digital processing and encoding devices. However, the same digital signal processing technologies that make high quality audio recordings feasible at a low cost also promote a widespread access to sophisticated tools intended for editing and tampering of digital audio.

The development of editing technologies reached levels that allow one to easily create a tampered version of an audio. For instance, users with little training can edit audio recordings including different events with such quality that may get unnoticed by an ordinary hearsay [1]. As a natural conclusion, the forensic examination of digital audio recordings is essential to identify the existence of such issues.

There are in the literature several schemes for tampering detection of audio recordings. For instance, in [2], the authors propose a splicing detection method by revealing abnormal differences in the local noise levels, based on an observed property that audio signals tend to have kurtosis close to a constant in the band-pass filtered domain. In [3], a statistical technique to model and estimate the amount of reverberation and background noise is proposed as a tool to identify tampers by a stability analysis.

Some techniques analyze effects produced by MP3 compression. In [4], the authors propose a way to expose forgeries identifying inconsistencies in quantization offsets features. In [5], the authors detect MP3 double compression extracting statistical features from the modified discrete cosine transform and applying a Support Vector machine (SVM) to these features.

Other techniques detect nonlinearities produced by abrupt transitions in audio signal using bispectral analysis [6] and higher order statistics to identify microphones and then reveal possible forgeries [7]. Also, in [8], an adaptive audio fingerprinting scheme is used to detect forgeries produced by the replication of short audio intervals inside the same recording.

Despite the large number of different approaches, the techniques that use the power grid signal interference component information, hereinafter referred to as Electric Network Frequency (ENF), are widely cited and used by the forensic community for forgeries detection [9]–[14]. In fact, the ENF signal is often found embedded in audio recordings from eavesdropping devices. The high availability associated with the well behaved characteristics make it an attractive feature under the forensic point of view, which explains the wide use of it.

In this paper, we propose the ESPRIT-Hilbert based tampering detection scheme with SVM classifier for audio forensic analysis, hereinafter referred to as SPHINS. It is a robust method for tampering detection, since it combines the ES-

Paulo Max G. I. Reis is with the National Institute of Criminalistics and Department of Electrical Engineering, University of Brasilia, Brazil (e-mail: paulo.pmgir@dpf.gov.br).

João Paulo C. L. da Costa is with the Department of Electrical Engineering, University of Brasilia, Brazil, Institute for Information Technology, Ilmenau University of Technology, Ilmenau, Germany, and Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany (e-mail: joaopaulo.dacosta@ene.unb.br)

Ricardo K. Miranda is with the Department of Electrical Engineering, University of Brasilia, Brazil, and Institute for Information Technology, Ilmenau University of Technology, Ilmenau, Germany (rickehrle@gmail.com).

Giovanni Del Galdo is with Institute for Information Technology, Ilmenau University of Technology, Ilmenau, Germany, and Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany (email:giovanni.delgaldo@iis.fraunhofer.de).

PRIT and Hilbert approaches for ENF estimation. In addition, SPHINS incorporates an outlier detector based on the sample kurtosis of the estimated ENF as input for a SVM classifier, outperforming the state-of-the-art ENF estimator approaches.

This paper is divided into seven sections including this introduction. In Section II, basic concepts on ENF signals are adressed and related works are reviewed. In Section III a model to characterize the ENF signal embedded in digital audio recording is presented. Section IV describes two different state-of-the-art ENF estimators: ESPRIT-based ENF Estimator (3E) and a Hilbert-based ENF Estimator (HEE). The ESPRIT-Hilbert based technique with SVM classifier is proposed in Section V. Section VI evaluates the method performance using the corpus Carioca 1 for different practical scenarios. The conclusions are then draw in Section VII. The main variables an parameters that appear in this paper are listed in Table I, while the acronyms used throughout the paper are listed in Table II.

TABLE I
MAIN VARIABLES AND PARAMETERS

Var./Par.	Description
$s_{\mathrm{ut}}(n)$	audio signal under test
$v(n)$	clean speech signal
$x(n)$	power grid interference signal
$e(n)$	additive noise signal
f_{nom}	power grid nominal frequency
$\omega(n)$	normalized ENF angular frequency
$f(n)$	ENF linear frequency
f_{s}	sampling rate
$\mathrm{BW}_{\mathrm{ENF}}$	ENF centered bandwidth
$\hat{x}(n)$	estimated power grid signal
$\hat{x}_{\mathrm{a}}(n)$	analytical approximation of $\hat{x}(n)$
$\hat{\omega}_{\mathrm{H}}(n)$	Hilbert-based ENF estimate
$\hat{\omega}_{\mathrm{E}}(n_b)$	Esprit-based ENF estimate
$s_{\mathrm{ds}}(n)$	down-sampled audio signal
$\hat{\mathrm{SNR}}_{\mathrm{ENF}}$	ENF signal-to-noise ratio
$\mathbf{F}$	feature vector
c_i	class index
$K(\cdot,\cdot)$	kernel function

TABLE II
ACRONYMS

Acr.	Description
ENF	Electrical Network Frequency
SVM	Support Vector Machine
HEE	Hilbert-based ENF Estimator
3E	Esprit-based ENF Estimator
EER	Equal Error Rate
FNR	False Negative Rate
FPR	False Positive Rate
OER	Overall Error Rate
DET	Detection Error Trade-off
VAD	Voice Activity Detector
SL	Saturation Level
BR	Bit Rate

II. RELATED WORKS

ENF signals are frequently found in audio recordings produced by telephone wiretapping, environmental eavesdropping devices, as well as in other types of audio signals with forensic interest. Even taps that use battery can present a considerable interference from the power grid. As an example, despite the efforts to produce an effective electrical insulation in telephone cables, it is common to see interference signals from the power grid in PSTN last mile.

The power grid signal is a standard signal with well-known characteristics. For instance, the nominal values of ENF are 50 Hz and 60 Hz depending on the place in question. European countries, Australia, and most of the countries in Asia and Africa use 50 Hz value. North and Central America countries use 60 Hz. Note that in South America, there are countries using 50 Hz and also other countries using 60 Hz. Japan uses both values as ENF nominal values [11]. Although the power grid signal is ideally a real sinusoidal signal that oscilates with a nominal frequency, the actual ENF signal presents variations in its instantaneous oscilation frequency due to the mismatch between energy supply and demand on the power grid. Indeed, these signals are quite well behaved, showing no abrupt frequency and phase changes over time. The good behavior of these signals is a mandatory feature to the correct operation of many electric and electronic equipments. Therefore, despite the existent variations, strict control mechanisms are used to maintain the frequency of power grid signal within very narrow limits.

The ENF signal eventually encountered in an audio recording is useful in forensic context since such signals can be used as a way to authenticate, identify the location of a recording and detect adulterations [9]. Since the ENF has a natural regularity, the existence of abrupt changes on the estimated ENF from an audio recording may indicate an adulteration. In fact, an insertion or deletion of audio segment in a recording has great chance to produce discontinuities in the instantaneous phase information and generate abnormal variations on the registered ENF. An automated approach is proposed by [10] to extract the ENF information by means of a quadratic interpolation of peak values in the Fourier spectrum. Extracted information can be compared with a stored database of regional ENF variations to authenticate an audio. However, this approach needs that regionals ENF databases are available for authentication purposes. In [11], a method is proposed based on detecting phase discontinuity in embedded ENF using a high-precision Fourier analysis to estimate the phase variations. If the variance of phase estimates exceeds an empirically defined threshold, there is an abnormal behavior typically derived from suppressions or insertions of audio segments.

In [12], the authors evaluate the performance of two different methods to estimate the ENF: MUltiple Signal Classification (MUSIC) [15] and Estimation of Signal Parameters via Rotational Invariance Techniques (ESPRIT) [16], [17]. Both schemes are compared with the approach used in [10] and ESPRIT proved to be more accurate.

More recently, [14] uses the Hilbert's analytic signal method for instantaneous frequency estimation to obtain the ENF information. Authors also establish an adaptive threshold for ENF excursion, based on a upper limit for normal variations observed in unedited signals to perform an automatic criteria

to classify tampered audio. For that purpose, public databases with audio speech recordings are used.

Despite the interesting results, both works in [14] and [11] assess the performance in terms of an equal error rate (EER), which is estimated from the number of misclassifications occurred at the same database used to define the thresholds, with no cross-validation approach employed. Also, both methods present a considerable degradation in low signal to noise ratio and high saturation levels of the registered signal.

This paper proposes a technique designated as SPHINS, aimed to reveal digital audio editions by means of a robust approach to analyze disruptions in power grid interference signal. Thereby, our ESPRIT-Hilbert based approach jointly exploits the Hilbert ENF estimates, which are more sensible to abrupt phase discontinuities, and the ESPRIT ENF estimates, which are more robust to noise. SPHINS uses a feature vector that summarizes the ENF disturbances based on the sample kurtosis of ENF estimates, applied into a SVM classifier to identify the presence of tampering.

III. Data Model

Let $s_{\rm ut}(n) \in \mathbb{R}$ be the audio signal under test which can be represented as the following superposition:

$$s_{\rm ut}(n) \;=\; v(n) + x(n) + e(n), \tag{1}$$

where $v(n)$ is the clean speech signal, $x(n)$ is the power grid interference signal and $e(n)$ is an additive noise portion. In the data model established, signals and noise are assumed uncorrelated with each other.

The power grid signal is ideally a real sinusoidal signal that oscilates with a nominal frequency $f_{\rm nom}$. However, there are variations in its instantaneous ENF due to mismatch between supply and demand on the grid. Despite the tight control mechanisms made by the power generation units, in practice it is impossible to instantaneously track and control all changes on the demand [18]. As a rule of thumb, the ENF is directly proportional to the power demand slightly varying around its nominal value. The limits of admitted frequency fluctuations depends on local regulations. In Brazilian power network, for instance, the ENF shall remain between 59.9 Hz and 60.1 Hz in steady state normal conditions of operation [19]. In general, an abrupt change on instantaneous frequency is not expected, which makes ENF signal interesting for forensic analysis.

This leads to the following model for the power grid signal:

$$x(n) \;=\; A(n)\cos\left(\theta_0 + \omega(n)\; n\right), \tag{2}$$

where θ_0 is the initial phase of the signal, and $\omega(n)$ is the electrical network normalized angular frequency, related to the ENF $f(n)$ by:

$$\omega(n) \;=\; \frac{2\pi f(n)}{f_{\rm s}}. \tag{3}$$

Our goal is given only $s_{\rm ut}(n)$ to detect a possible tampering by exploiting the variations of the estimate $\hat{\omega}(n)$.

IV. ENF Estimation Approaches

To ENF extraction $s_{\rm ut}(n)$ must be processed by a very sharp bandpass filter with bandwidth $\mathrm{BW}_{\rm ENF}$ centered in $f_{\rm nom}$, producing power grid signal estimate $\hat{x}(n)$:

$$\begin{aligned} \hat{x}(n) \;&=\; h_{\rm bp}(n) * s_{\rm ut}(n), \\ \hat{x}(n) \;&=\; h_{\rm bp}(n) * x(n) + h_{\rm bp}(n) * [v(n) + e(n)], \end{aligned} \tag{4}$$

where $h_{\rm bp}(n)$ is an impulse response of the bandpass filter and the operator $*$ denotes a linear convolution.

The bandwidth $\mathrm{BW}_{\rm ENF}$ must be as narrow as possible to reject all the signals other than $x(n)$, and sufficiently large to cover all ENF excursion, such that we can write:

$$h_{\rm bp}(n) * x(n) \;\approx\; x(n), \tag{5}$$

Considering audio recordings obtained by a PSTN wiretapping, there is almost no spectral superposition between power grid and voice signals, such that we can write:

$$\hat{x}(n) \approx x(n) + z(n), \tag{6}$$

where $z(n)$ corresponds to the sum of the filtered noise term and a bandpass filtered speech residual term.

In this section, we give an overview of two approaches for ENF estimation from $\hat{x}(n)$. In Subsection IV-A, we present the Hilbert-based ENF estimator and, in Subsection IV-B, the ESPRIT based ENF estimator.

A. Hilbert-based ENF Estimator (HEE)

The HEE is a Hilbert transform based approach to estimate the ENF using instantaneous phase information of the Hilbert's analytical approximation $\hat{x}_a(n) \in \mathbb{C}$ of a real valued estimate $\hat{x}(n)$ [14]:

$$\hat{x}_{\rm a}(n) \;=\; \hat{x}(n) + j\mathcal{H}\left\{\hat{x}(n)\right\}, \tag{7}$$

where the operator $\mathcal{H}\{\}$ denotes the Discrete Hilbert Transform and $j = \sqrt{-1}$.

Replacing (2) in (6), we have:

$$\hat{x}_{\rm a}(n) \;=\; A(n)\exp j\left(\theta_0 + \omega(n)\; n\right) + z_{\rm a}(n), \tag{8}$$

where $z_{\rm a}(n)$ is the analytical representation of the noise term $z(n)$.

The ENF estimate $\hat{\omega}_H(n)$ obtained sample-by-sample is given by the first order derivative[1] of the instantaneous phase estimate:

$$\hat{\omega}_{\rm H}(n) \;=\; \dot{\hat{\theta}}(n), \tag{9}$$

where $\dot{\hat{\theta}}(n)$ is the first derivative of:

$$\hat{\theta}(n) \;=\; \angle\hat{x}_{\rm a}, \tag{10}$$

[1]The fisrt derivative is here defined as $\dot{\hat{\theta}}(n) = \hat{\theta}(n) - \hat{\theta}(n-1)$.

where $\angle\hat{x}_{\rm a}$ corresponds to the argument of $\hat{x}_{\rm a}$, given by:

$$\angle\hat{x}_{\rm a} \;=\; \arctan\left(\frac{\Im(\hat{x}_{\rm a})}{\Re(\hat{x}_{\rm a})}\right), \tag{11}$$

where $\Im(\hat{x}_{\rm a})$ and $\Re(\hat{x}_{\rm a})$ are the imaginary and real parts of $\hat{x}_{\rm a}$, respectively.

If $\hat{\theta}(n)$ varies slowly compared to the sampling rate, a reasonable approximation to the first derivative is given by [14]:

$$\hat{\omega}_{\rm H}(n) = \angle\left(\hat{x}_{\rm a}(n)\hat{x}^{*}_{\rm a}(n-1)\right), \tag{12}$$

where the operator * is used to denote the complex conjugate.

B. ESPRIT-based ENF Estimator (3E)

The 3E is a subspace-based approach to compute ENF using the rotational property between staggered subspaces to estimate the frequency [17].

Considering (8), let $\mathbf{X}$ be an $N \times M$ data matrix such that:

$$\mathbf{X} \;=\; [\hat{\mathbf{x}}_{\mathbf{a}}(0) \quad \hat{\mathbf{x}}_{\mathbf{a}}(1) \quad \ldots \quad \hat{\mathbf{x}}_{\mathbf{a}}(N-2) \quad \hat{\mathbf{x}}_{\mathbf{a}}(N-1)]^{\rm T}, \tag{13}$$

where $\hat{\mathbf{x}}_{\mathbf{a}}(n) = [\hat{x}_{\rm a}(n)\ \hat{x}_{\rm a}(n+1)\ \ldots\ \ \hat{x}_{\rm a}(n+M-1)]^{\rm T}$, and $()^{\rm T}$ denotes the transpose operation.

The matrix $\mathbf{X}$ can be decomposed by applying a Singular Value Decomposition (SVD):

$$\mathbf{X} \;=\; \mathbf{LSU}^{\rm H}, \tag{14}$$

where $\mathbf{L}$ is an $N\times N$ matrix containing the left singular vectors of $\mathbf{X}$, $\mathbf{S}$ is an $N \times M$ matrix containing the singular values of $\mathbf{X}$ in its diagonal, $\mathbf{U}$ is an $M \times M$ matrix containing the right singular vectors of $\mathbf{X}$, and $()^{\rm H}$ corresponds to the Hermitian operator.

The matrix $\mathbf{U}$ can be further decomposed as $\mathbf{U} = [\mathbf{u}_s|\mathbf{U}_{noise}]$, where $\mathbf{u}_s$ is the $M\times 1$ signal vector formed by the singular vector corresponding to the largest singular values of $\mathbf{X}$. The remaining singular vectors form the $M\times(M-1)$ noise matrix $\mathbf{U}_{noise}$. Note that the model order of matrix $\mathbf{X}$ is equal to one, since the signal has only one frequency component.

Note also that the signal subspace is rotational invariant. Let us consider that the amplitude and the angular frequency in (8) are almost constant among all samples in $\mathbf{X}$. Thus, we can argue that the following $\mathbf{d}$ vector belongs to the signal subspace:

$$\mathbf{d} \;=\; \left[1\ e^{j\omega}\ e^{2j\omega}\ \cdots\ e^{(M-1)j\omega}\right]^{\rm T} \tag{15}$$

Considering that $\mathbf{d_u}$ and $\mathbf{d_d}$ are the vectors formed by the first and the last $M-1$ elements of $\mathbf{d}$, it holds that:

$$\mathbf{d_u}e^{j\omega} \;=\; \mathbf{d_d}, \tag{16}$$

satisfying the rotational invariance property.

Let $\mathbf{u_u}$ and $\mathbf{u_d}$ be the vectors formed by the first and the last $M-1$ elements of $\mathbf{u_s}$. Since $\mathbf{u}_s$ spans the same signal subspace of $\mathbf{d}$, the rotational invariance property holds and allows us to write:

$$\mathbf{u_u}\phi \;=\; \mathbf{u_d}, \tag{17}$$

where ϕ is a complex exponential whose argument corresponds to the desired angular frequency ω. Solving (17) leads to the ESPRIT estimated electric network angular frequency $\hat{\omega}_{\rm E}$:

$$\hat{\omega}_{\rm E} \;=\; \angle\frac{\mathbf{u_u}^{\rm H}\mathbf{u_d}}{\mathbf{u_u}^{\rm H}\mathbf{u_u}}. \tag{18}$$

Unlike Hilbert approach, ESPRIT method results in a fixed parameter that is an optimal estimate of ENF in a least square sense for the given input data. To obtain a time varying estimate of ENF, one can divide the input signal $\hat{x}_{\rm a}(n)$ into N_b short term blocks with N samples per block and $N-M$ overlapping samples. The ESPRIT method is used on each block to estimate the angular frequency. As a result we obtain a sequence of samples that represents the time varying ENF signal $\hat{\omega}_{\rm E}(n_b)$ for each consecutive block.

V. Proposed ESPRIT-Hilbert based tampering detection scheme with SVM classifier

Similarly to the state-of-the-art schemes, the proposed method requires the validity of certain assumptions. First, the power grid interference is the most intense signal in the narrow vicinity of nominal ENF, ensuring a signal to noise ratio (SNR) in this spectral neighborhood that allows a reliable estimate of the ENF. With further degradation of the SNR, there is a deterioration in the quality of ENF estimate, resulting in a worse performance in tampering detection. Moreover, no significant level of nonlinear distortions is assumed. Finally, insertions or suppressions of audio segments are made in non-active voice instants.

In this section the proposed ESPRIT-Hilbert based tampering detection scheme with SVM classifier (SPHINS) is discussed. The block diagram of Fig. 1 shows the proposed scheme divided in three blocks. In Subsection V-A, Block 1 in Fig. 1 about preprocessing procedures is described. In Subsection V-B, Block 2 in Fig. 1 is discussed exploring the ESPRIT-Hilbert feature extraction. Finally, the classification step of Block 3 in Fig. 1 is detailed in Subsection V-C.

A. Preprocessing stage

Let $s_{\rm ut}(n)$ in (1) be the questioned audio signal and $f_{\rm nom}$ the nominal ENF in Hz. Before the feature extraction and classification task, $s_{\rm ut}(n)$ passes through a preprocessing stage divided in the steps illustrated in the block diagram shown in Fig. 2.

In Block 1 of Fig. 2, signal $s_{\rm ut}(n)$ is down-sampled to a sampling rate $f_{\rm s} = 20f_{\rm nom}$, ensuring an exact number of 20 samples per nominal ENF period. If the original sampling rate is not an integer multiple of $f_{\rm s} = 20f_{\rm nom}$, we first perform an interpolation, increasing the original sampling rate by a factor of $f_{\rm s}/\alpha$, where α is the great common divider between the original sampling rate and the target sampling rate $f_{\rm s}$. After that, a decimation procedure is made, with a proper anti-aliasing filtering. The down-sampled signal is called $s_{\rm ds}(n)$.

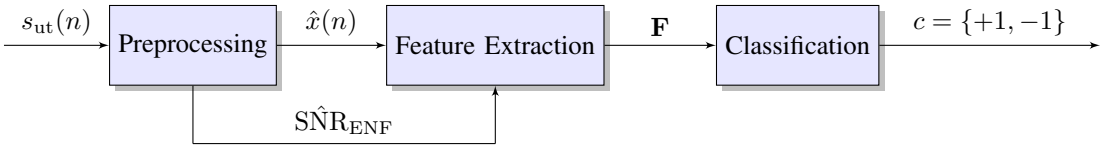

Fig. 1. Block diagram that illustrates the three stages of the proposed method

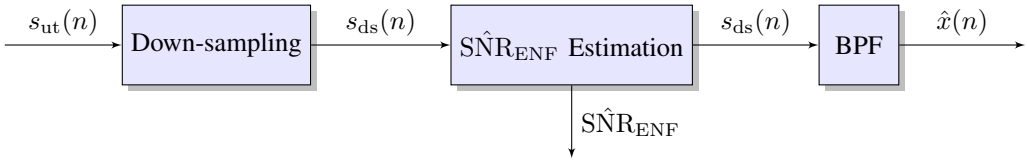

Fig. 2. Block diagram of preprocessing stage

$\hat{\text{SNR}}_{\text{ENF}}$ is calculated via a rough estimate of the overall ENF SNR in a narrow spectral vicinity of f_{nom}. This estimate is used as an element of the feature vector formed in Subsection V-B. The $\hat{\text{SNR}}_{\text{ENF}}$ estimate can be obtained considering that ENF is the most intense signal confined in a narrow vicinity of f_{nom}, and that the noise in this narrow spectral region is well approximated to a white uncorrelated noise.

Then, in order to compute $\hat{\text{SNR}}_{\text{ENF}}$, we first calculate the power spectral density estimate $\hat{P}_{\text{ds}}(k)$ of $s_{\text{ds}}(n)$ in dB, where $k = 0, 1, 2..., N_{FFT}/2$ and N_{FFT} is the number of points used in FFT algorithm. For that purpose, we use the Fast Fourier Transform (FFT) based method proposed in [20]. Fig. 3 illustrates $\hat{P}_{\text{ds}}(k)$ in a limited neighborhood of f_{nom} for a typical wiretapped audio recording.

To compute $\hat{\text{SNR}}_{\text{ENF}}$, we estimate the noise floor in a narrow spectral vicinity in $\hat{P}_{\text{ds}}(k)$ which corresponds to a bandwidth of 3BW_{ENF} centered in the f_{nom} exemplified as 60 Hz. Remembering that ENF signal is modeled as a real sinusoidal signal with a frequency excursion confined in a narrow bandwidth BW_{ENF} centered in f_{nom}, we also estimate the peak value of $\hat{P}_{\text{ds}}(k)$ in this narrow vicinity. The estimated $\hat{\text{SNR}}_{\text{ENF}}$ is then calculated as the difference, in dB, between the peak power density value and the noise floor.

Formally, let $\hat{P}_{\text{ENF}}$ and $\hat{P}_{\text{noise}}$ be:

$$\hat{P}_{\text{noise}} = 10\log_{10}\left(\frac{1}{|\mathbf{\Omega_2}-\mathbf{\Omega_1}|}\sum_{k\in(\mathbf{\Omega_2}-\mathbf{\Omega_1})} 10^{\hat{P}_{\text{ds}}(k)/10}\right),$$

$$\hat{P}_{\text{ENF}} = 10\log_{10}\left(10^{\max[\hat{P}_{\text{ds}}(k)]/10} - 10^{\hat{P}_{\text{noise}}/10}\right),\ k \in \mathbf{\Omega_1}, \tag{19}$$

where $\mathbf{\Omega_1}$ and $\mathbf{\Omega_2}$ are the subsets:

$$\mathbf{\Omega_1}: \frac{N_{\text{FFT}}}{fs}(f_{\text{nom}} - \frac{\text{BW}_{\text{ENF}}}{2}) \le k \le \frac{N_{\text{FFT}}}{fs}(f_{\text{nom}} + \frac{\text{BW}_{\text{ENF}}}{2}),$$

$$\mathbf{\Omega_2}: \frac{N_{\text{FFT}}}{fs}(f_{\text{nom}} - \frac{3\text{BW}_{\text{ENF}}}{2}) \le k \le \frac{N_{\text{FFT}}}{fs}(f_{\text{nom}} + \frac{3\text{BW}_{\text{ENF}}}{2}). \tag{20}$$

In Fig. 3, the shaded area corresponds to the $\mathbf{\Omega_2} - \mathbf{\Omega_1}$ and the bright area to $\mathbf{\Omega_1}$ for a $\text{BW}_{\text{ENF}} = 0.8$ Hz and

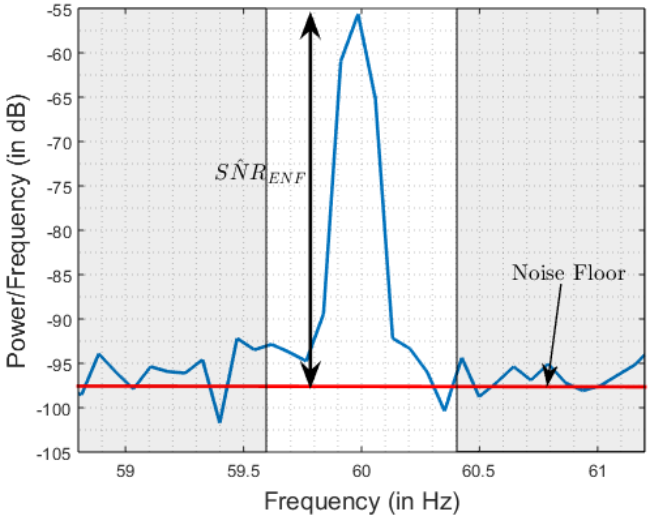

Fig. 3. Power Espectral Density estimate $\hat{P}_{\text{ds}}(k)$ of a typical wiretapped audio recording with embedded ENF signal in a narrow vicinity of the ENF nominal frequency. Shaded area corresponds to $\mathbf{\Omega_2} - \mathbf{\Omega_1}$ and non-shaded area to $\mathbf{\Omega_1}$

$f_{\text{nom}} = 60$ Hz. To finally compute SNR_{ENF} we simply calculate a difference between terms in (19):

$$\hat{\text{SNR}}_{\text{ENF}} = \hat{P}_{\text{ENF}} - \hat{P}_{\text{noise}}. \tag{21}$$

Finally, $s_{\text{ds}}(n)$ is filtered with a sharp zero-phase bandpass FIR filter (BPF) around f_{nom}, with a bandwidth BW_{ENF} and using L coefficients. This bandpass filtered signal corresponds to the $\hat{x}(n)$ in (6).

B. Proposed ESPRIT-Hilbert Feature Extraction

One of the main contributions of this article is the proposal of a new feature vector to characterize a ENF disturbance produced by insertions and deletions in audio recordings, and the incorporation of machine learning techniques to solve the detection of disturbances. After the preprocessing stage, two ENF estimators are used, namely, HEE and 3E, resulting in two different estimates $\hat{\omega}_{\text{H}}(n)$ and $\hat{\omega}_{\text{E}}(n_b)$, respectively.

In order to permit a direct comparison between the two estimates, we consider the HEE estimates block by block, not sample by sample, using their values in the center of each block, as following:

$$\hat{\omega}_{\mathrm{H_b}}(n_b) = \hat{\omega}_{\mathrm{H}}\left(\left\lceil n_b M + \frac{N}{2} - 1 \right\rceil\right), \tag{22}$$

where $\lceil\cdot\rceil$ is the ceiling operator, that is, it returns the smallest integer greater than or equal to the operand.

Fig. 4 illustrates the results of two estimators when applied to a 28 seconds audio signal with segment suppression. As shown in Fig. 4, both estimators produce very similar results capturing together the ENF variations over the time, except in the suppression point, where a significant difference can be observed. In fact, HEE is more sensible to abrupt phase discontinuities than 3E, whose parametrized model tends to force a perfectly sinusoidal model estimation for each block. Considering these aspects, we propose a feature aimed to exploit jointly the Hilbert ENF estimates, which are more sensible to abrupt phase discontinuities, and the ESPRIT ENF estimates, which are more robust to noise.

To classify automatically an audio signal as edited or unedited, a feature that carries information about abnormal variations in ENF estimate is needed. For this purpose, the sample kurtosis of the ENF estimates is used.

Unlike variance, kurtosis is a scale independent parameter that measures the tailedness of a distribution. The presence of symmetrical or asymmetrical outliers in data distribution has the property of increasing the original kurtosis [21]. Therefore, kurtosis measures the influence of extreme values in overall variance of the data.

The sample kurtosis of $\hat{\omega} = \{\hat{\omega}(n_b) : n_b = 1, 2 \ldots N_b\}$ can be defined as the ratio between fourth sample central moment and squared sample variance:

$$\kappa(\hat{\omega}) = \frac{1/N_b \sum_{i=1}^{N_b} (\hat{\omega}(n_b) - \bar{\hat{\omega}})^4}{(1/N_b \sum_{i=1}^{N_b} (\hat{\omega}(n_b) - \bar{\hat{\omega}})^2)^2}, \tag{23}$$

where $\bar{\hat{\omega}} = 1/N_b \sum_{i=1}^{N_b} \hat{\omega}(n_b)$.

Using the kurtosis as an outlier measure for both HEE and 3E estimates, the proposed feature vector is defined as:

$$\mathbf{F} = \left[\kappa(\hat{\omega}_{\mathrm{H_b}})\ \kappa(\hat{\omega}_{\mathrm{E}})\ \mathrm{SNR}_{\mathrm{ENF}}\right], \tag{24}$$

where $\kappa(\hat{\omega}_{\mathrm{H_b}})$ and $\kappa(\hat{\omega}_{\mathrm{E}})$ are the kurtosis of ENF estimates, and $\mathbf{F} \in \mathbb{R}^3$ is the feature vector that summarizes anomalous ENF variations for an arbitrary audio recording. Since the HEE and 3E estimates have different robustness to noise, we included $\mathrm{SNR}_{\mathrm{ENF}}$ as an element in feature vector. This allows the SVM classifier learning an appropriate decision surface in the training stage, jointly considering the $\mathrm{SNR}_{\mathrm{ENF}}$ value with the kurtosis of the ENF estimates.

C. Classification Stage

To classify an audio as tampered, we train SVM classifier [22] with a training dataset $(\mathbf{F}_i, c_i)$ of an equal number of edited and unedited audio recordings from a known database, where $\mathbf{F}_i$ is the feature vector in (24), and $c_i = \{+1, -1\}$ is the class index corresponding to the i-th audio recording, for $i = 1, \ldots, N_{ar}$, with N_{ar} being the total number of audio recordings in the database.

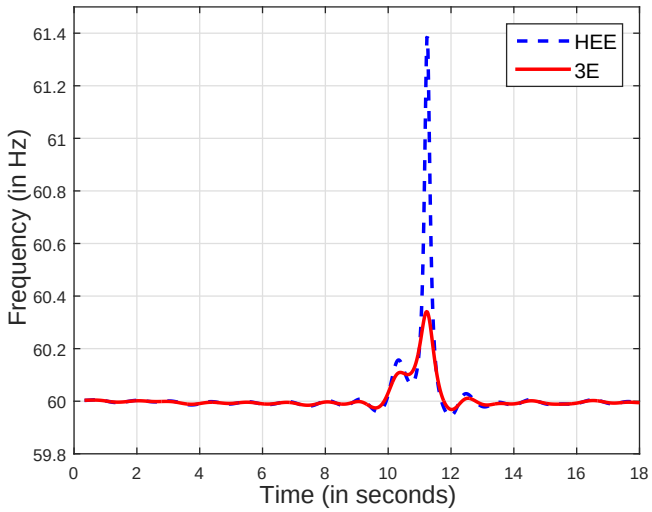

Fig. 4. Hilbert vs. ESPRIT ENF estimators: ENF estimates superposed

More specifically, we train a classifier finding the discriminant function $g(\cdot)$ such that:

$$g(\mathbf{F}_i) = \begin{cases} \geq t, & c_i = +1 \\ < t, & c_i = -1. \end{cases} \tag{25}$$

where t is a threshold.

In the training phase, SVM computes the maximum margin separation hyperplane in the feature space which separates data from two classes with the maximum distance. This hyperplane is fully determined by the unitary vector $\mathbf{a}$ normal to the hyperplane and the bias b which corresponds to the perpendicular distance from hyperplane to the origin, such that the training phase consists in determining these two parameters [22], such that $g(\mathbf{F}_i) = \mathbf{a}^{\mathrm{T}}\mathbf{F}_i + b$, where the relation $\mathbf{a}^{\mathrm{T}}\mathbf{F}_i + b = 0$ holds for points $\mathbf{F}_i$ that belongs to this maximum margin hyperplane.

For nonlinear separable data, the maximum margin separation surface can be found considering new feature vectors $\eta(\mathbf{F}_i)$ mapped to a higher dimensional space in which they are linear separable, where the function $\eta(\cdot) : \mathbb{R}^{d_1} \rightarrow \mathbb{R}^{d_2}$, with $d_1 < d_2$, is the function applied to map the original feature vectors to a higher dimensional. In this higher dimensional space the classifier discriminant function $g(\cdot)$ is such that

$$g(\mathbf{F}_i) = \mathbf{a}^{\mathrm{T}}\eta(\mathbf{F}_i) + b, \tag{26}$$

and the relation $g(\mathbf{F}_i) = 0$ holds for points $\eta(\mathbf{F}_i)$ that belongs to the maximum margin surface.

In practice, SVM uses only the inner products $\langle \eta(\mathbf{F}_i), \eta(\mathbf{F}_j) \rangle$ of the mapped feature vectors to find and evaluate the discriminant function $g(\cdot)$. Instead of performing the expensive mapping of each feature vector to a higher dimensional space, SVM computes these inner products of mapped feature vectors directly from the original space by means of the kernel trick, in which:

$$K(\mathbf{F}_i, \mathbf{F}_j) = \langle \eta(\mathbf{F}_i), \eta(\mathbf{F}_j) \rangle \tag{27}$$

where $K(\cdot,\cdot)$ are the kernel function. For more detailed information, the reader is referred to [22].

The proposed SPHINS method uses a Gaussian radial basis function as the kernel function, defined as the following:

$$K(\mathbf{F}_i, \mathbf{F}_j) = \exp\left(-\frac{||\mathbf{F}_i - \mathbf{F}_j||^2}{2\sigma^2}\right). \tag{28}$$

We choose a Gaussian kernel function due to its simplicity, mathematical tractability, numerical stability, and due to its advantages for being an universal kernel, i. e., a kernel function able to find Hilbert spaces with universal approximating capability, usually giving reasonable results [23]. The parameter σ is an empirically determined optimal parameter that controls the bias-variance trade-off [22]. Typically the optimal σ is determined in cross validation tests.

An audio recording that results in a feature $\mathbf{F_i}$ is classified as of a class $c_i = +1$ if the calculated score $g(\mathbf{F}_i) \geq t$, and as $c_i = -1$, otherwise. In this work we set the decision threshold t as the point of operation to produce an EER in the training set, that is to produce the same number of false positive and false negative classifications in the training data.

VI. Experiments ans Results

In order to evaluate the performance of the proposed technique, known controlled edited and unedited audio signals are used. For that, a public corpus named Carioca 1 [11] is employed. The corpus is the same used in [11] and [14], allowing some direct comparison with previous related works, and can be obtained from the website http://lps.lncc.br/index.php/demonstracoes/tifs2014.

The Carioca 1 database has 100 unedited authorized audio recordings of phone calls, being 50 of them from male speakers and 50 from female speakers. Moreover, Carioca 1 corpus has 100 edited versions of the same phone calls, being 50 edited by one segment suppression and the other 50 by one segment insertion, all of them equally distributed in male and female speaker recordings. The insertions and suppressions are such that the related phase discontinuities are normally distributed between +180 degrees and -180 degrees [11]. The phone call recordings are sampled at 44.1 kHz with 16-bit quantization and coded with lossless linear Pulse Code Modulation. The duration varies from 19 s to 35 s and there is almost no interference between ENF and voice signals due to PSTN bandwidth.

In our experiments, we consider that $f_{\text{nom}} = 60$ Hz and $f_s = 1200$ Hz, ensuring a exact number of 20 samples per nominal ENF period. As in [11], we use $L = 10.000$ for sharp zero-phase bandpass FIR filter, and we set $N_{\text{FFT}} = 2^{14}$ in the $\hat{\text{SNR}}_{\text{ENF}}$ estimation to ensure a high frequency resolution in the power spectral density estimate $\hat{P}_{\text{ds}}(k)$. We use $\text{BW}_{\text{ENF}} = 0.8$ HZ, since the bandwidth must be as narrow as possible to reject noise and other spurious signals other than ENF, but sufficiently large to cover ENF excursion. In Subsection VI-A, we show SPHINS robustness over some different BW_{ENF} values in a critical SNR condition. We also use $\sigma = 3.5$, chosen empirically in a critical saturation scenario as described in Subsection VI-B.

The Carioca 1 database is used to train SVM classifier. After determining the nonlinear discriminant function (25), the audio recordings are classified for a wide range of thresholds t and the correspondent values of false positive rate (FPR) and a false negative rate (FNR) are computed, allowing to construct the detection error trade-off (DET) curve [24]. The EER is the point in DET curve in which FPR=FNR and the value of t that corresponds to EER is chosen as the decision threshold in (25) for the classification task.

Although the EER be a useful parameter to characterize and compare different system performances, its inherent re-substitution procedure gives an optimistic error rate for the classification task. For a more realistic evaluation, a cross validation procedure is made with a 10-fold strategy where the entire database is equally divided in ten subgroups in which nine of them is used to train the classifier. The EER point of operation is used to obtain the classifier threshold t to test the remaining subgroup. This procedure is repeated ten times changing to a different test subgroup in each iteration and counting the misclassifications.

To evaluate the influence of the 3E block size in our algorithm, we assess the method performance among several block sizes by conducting cross-validation tests for sets of $N = \{100, 200, 300, 400, 500, 600, 700, 800\}$ samples. As shown in Fig 5, there is a region of N values between 100 and 500 samples where the performance of SPHINS remain stable for the considered database, resulting in a overall error rate (OER) of 4.5 %, which corresponds to 95.5 % of accuracy, with 1 % of FPR and 8 % of FNR. Note that in the cross-validation procedure, different FPR and FNR are obtained as the EER point of operation is tuned to resubstitution tests. In all other tests we use $N = 200$ samples.

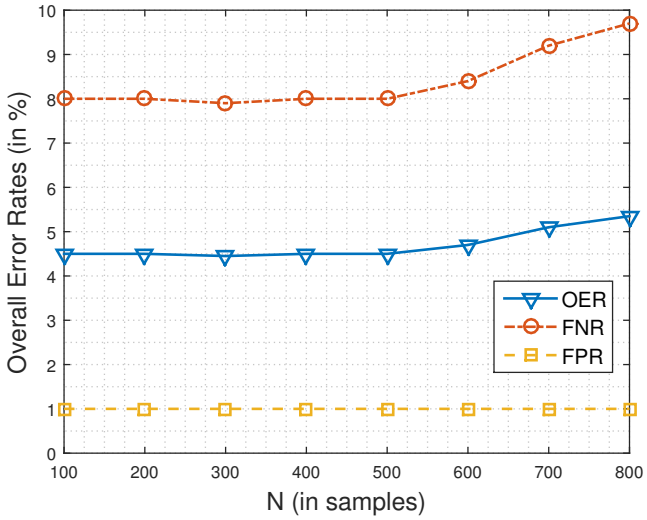

Fig. 5. SPHINS overall error rates with Carioca 1 database over different values of N.

Further, to compare with previous works we evaluate SPHINS DET curve as shown in Fig. 6. SPHINS attained an

EER of 4 % outperforming the EER of 7 % of method in [11] and with the same performance of [14] for the clean database.

The robustness of the proposed scheme is evaluated by experiments made simulating unfavorable conditions. In Subsection VI-A the proposed SPHINS method is evaluated in different conditions of SNR. In Subsection VI-B, the technique is evaluated at different levels of digital saturation. Finally, in Subsection VI-C, SPHINS is evaluated at different MP3 compression levels.

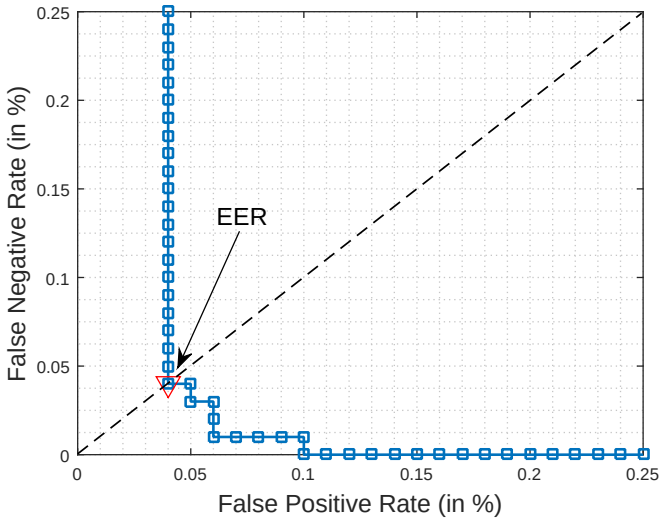

Fig. 6. DET curve showing FPR versus FNR for the Carioca 1 database. The 4 % EER is marked .

TABLE III
EQUAL ERROR RATES (EER) WITH CARIOCA 1 DATABASE DEGRADED WITH UNCORRELATED WHITE NOISE

SNR(dB)	EER-proposed(%)	[14]-EER(%)
30	4	4
25	5	5.4
20	7	12.3
15	10	24.7
10	21	38.7
5	21	45

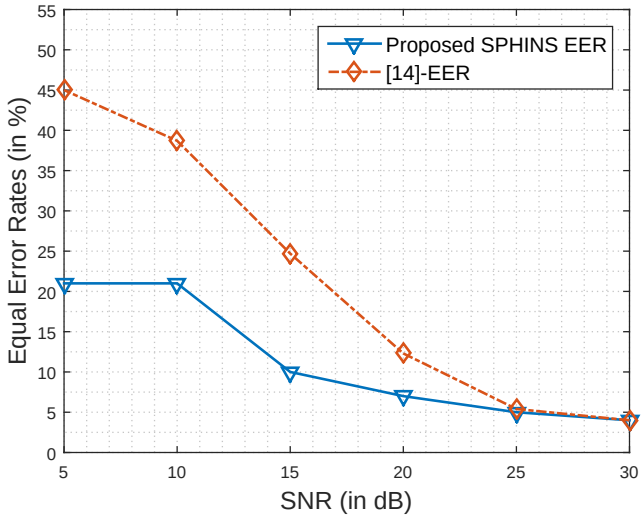

Fig. 7. Proposed Equal Error Rates with white Gaussian noise corrupted databases vs. [14] EER

A. Results Under Different SNR

In order to asses the robustness of the method against noise degradation, experiments corrupting the Carioca 1 audio database with white Gaussian noise in a controlled manner are performed. With the same approach used in [14], the original SNR of each audio signal in the corpus is measured using the same voice activity detector (VAD) in [14] to separate speech signal from background noise. An extra background noise is added to each audio to ensure a prescribed SNR whenever it has an original SNR higher than the target SNR. The set of prescribed SNR, in dB, varies from 5 to 30, with a 5 dB step. To ensure statistical variability, this task is made 10 times to produce 10 noisy versions of Carioca 1 database for each prescribed SNR. We then apply the proposed method in all 10 noisy versions of the database for each prescribed SNR, extracting the ESPRIT-Hilbert based feature vectors, training an SVM classifier and computing the EER. Finally, for each SNR target, the final EER is the mean of EER values computed in the correspondent noisy versions.

Fig 7 shows the EER obtained, comparing the proposed SPHINS with state-of-the-art approaches. According to Table III and Fig. 7, when applied to Carioca 1 database, SPHINS achieves a better performance than [14] in noise degraded scenarios.

For a more realistic evaluation, we perform a 10-fold cross validation avoiding using the same audio in training (picked from the clean database) and tests (picked from the noisy versions of Carioca 1), as shown in Table IV and Fig. 8. The same cross validation tests are not found in related work for appropriate comparison [11] [14].

TABLE IV
SPHINS CROSS-VALIDATED ERROR RATES WITH CARIOCA 1 DATABASE DEGRADED WITH UNCORRELATED WHITE GAUSSIAN NOISE

SNR(dB)	OER(%)	FPR(%)	FNR(%)
30	4.5	1	8
25	4.6	1	8.2
20	5.85	0.1	11.6
15	11.75	2.7	20.8
10	23.6	20.5	26.7
5	34.05	45.7	22.4

We also evaluate SPHINS performance for different values of $\mathrm{BW_{ENF}}$, varying from 0.4 Hz to 1.4 Hz, for a prescribed SNR of 10 dB. As can be seen in Fig. 9, there is a considerable performance degradation as $\mathrm{BW_{ENF}}$ grows up. The better performance is achieved for the smallest value of $\mathrm{BW_{ENF}} = 0.4$ Hz. However, the use of a very narrow bandwidth can cause a significant performance loss in cases of networks that present high ENF fluctuations. To prevent bad performances in this situations, SPHINS uses a slightly larger $\mathrm{BW_{ENF}} = 0.8$ Hz.

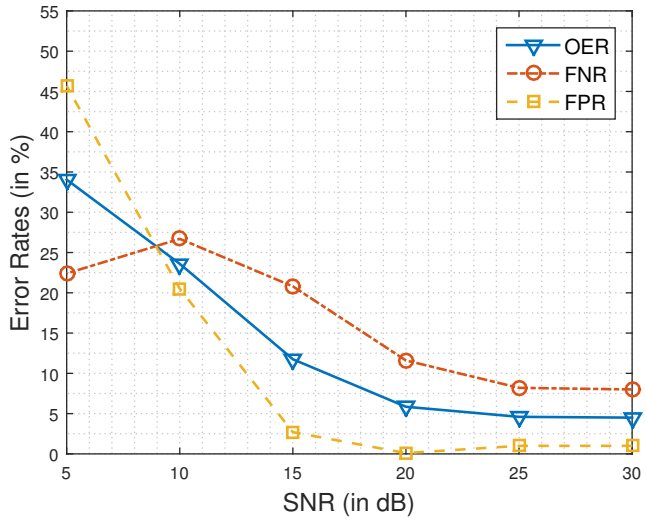

Fig. 8. SPHINS Cross-validated Error Rates With Carioca 1 Database Degraded With Uncorrelated White Gaussian Noise

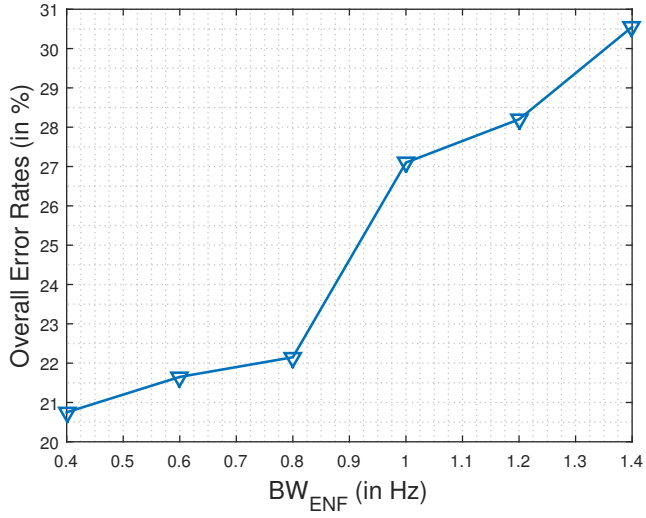

Fig. 9. SPHINS overall error rates with Carioca 1 database degraded to a prescribed SNR=10 dB, with uncorrelated white gaussian noise, over different values of $\mathrm{BW_{ENF}}$

B. Results Under Different Saturation Levels

A further source of deterioration with practical interest consists in nonlinearities caused by the audio signal clipping. To assess the effect of this degradation in the detection performance, the audios in Carioca 1 are subjected to different saturation levels (SL). Using the same VAD approach employed in [14], a percentage of active voice samples, which corresponds to a prescribed SL, is clipped to a maximum value. The set of prescribed saturation levels used is $\mathrm{SL} = [0,\ 0.2,\ 0.5,\ 1,\ 2,\ 4]$.

For each prescribed SL, the proposed method is applied computing the EER. According to Table V and Fig. 10, when applied to Carioca 1 database, the proposed SPHINS method outperforms [14] in digital saturation scenarios.

We also perform 10-fold cross validation tests for a more realistic evaluation of the proposed technique. As shown in Table VI and Fig. 11, the results are similar to those in Table V and Fig. 10.

TABLE V
EQUAL ERROR RATES (EER) WITH CARIOCA 1 DATABASE DEGRADED WITH DIFFERENT SATURATION LEVELS

SL(%)	EER-proposed(%)	[14]-EER(%)
0	4	4
0.2	8	12
0.5	8	12
1	10	13
2	11	14
4	14	18

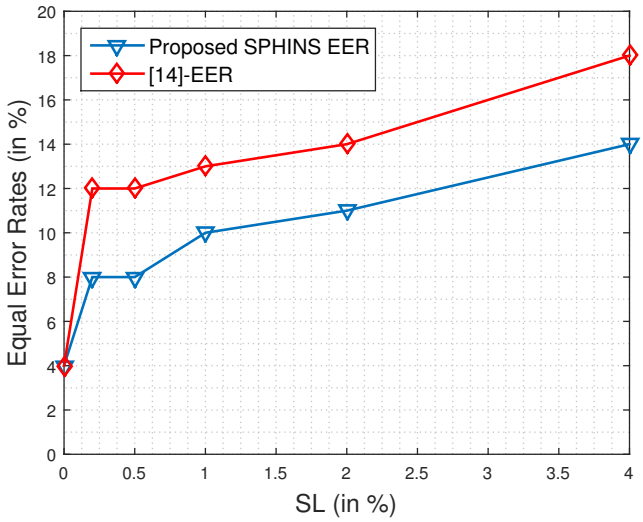

Fig. 10. Equal Error Rates (EER) with saturated databases vs. [14] EER

TABLE VI
SPHINS CROSS-VALIDATED ERROR RATES WITH CARIOCA 1 DATABASE DEGRADED WITH DIFFERENT SATURATION LEVELS

SL(%)	OER(%)	FPR(%)	FNR(%)
0	4.5	1.0	8
0.2	7.2	4.9	9.5
0.5	9.6	9.1	10
1	12.2	13.4	10.9
2	11.9	13.2	10.6
4	12.6	13.1	12.0

To asses the influence of σ parameter in high saturation scenarios, cross-validation tests are made for different values of σ between 2.1 and 4.6 for $\mathrm{SL} = 4\ \%$. Fig. 12 shows that the best performance is achieved for sigma values between 2.5 and 3.5 resulting in OER of 12.5 %. As expected, too low σ values lead to overfitting, resulting in a high overall error rate in cross validation tests. In all other tests we use $\sigma = 3.5$.

C. Results Under Different MP3 Compression Levels

A last set of tests made relate to the performance evaluation of the method in case of MP3 compressed audio signals. In order to asses the method for such type of signals, the

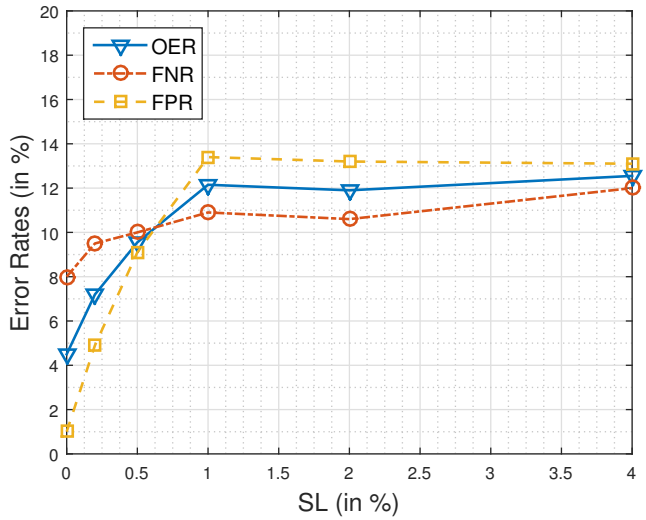

Fig. 11. SPHINS Cross-validated Error Rates With Carioca 1 Database Degraded With Different Saturation Levels

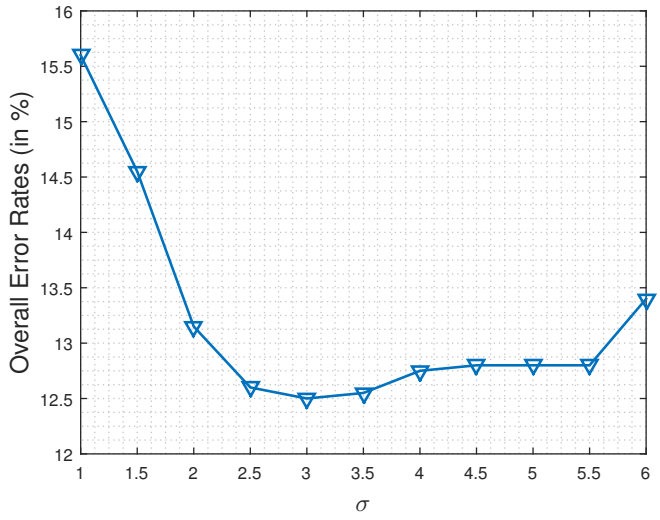

Fig. 12. SPHINS overall error rates with Carioca 1 database over different values of σ.

audio files in Carioca 1 database are subjected to several MP3 compression bit rates (BR). More specifically, we recode the entire database using a sampling rate of 22,050 Hz, and a set of prescribed bit rates BR $= [16, 32, 64, 128]$ kbps.

The proposed SPHINS is applied for each compression bit rate and the corresponding EER is then computed. According to Table VII and Fig. 13, there is a similar EER performance of the proposed method when applied to MP3 compressed audio files, with a minor degradation for low MP3 bit rates.

In addition, we perform 10-fold cross validation tests to evaluate the proposed technique under different MP3 compression bit rates in order to obtain a more realistic performance assessment. The results in Table VIII and Fig. 14 show that there is a performance degradation for highly compressed mp3 files.

TABLE VII
EQUAL ERROR RATES (EER) WITH CARIOCA 1 DATABASE RECODED WITH DIFFERENT MP3 COMPRESSION BIT RATES

BR(kbps)	EER-proposed(%)
16	4.5
32	5
64	4.5
128	4

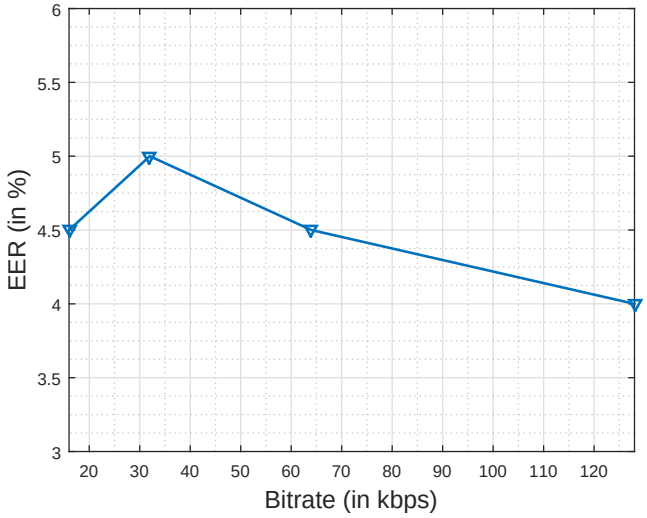

Fig. 13. Equal Error Rates (EER) with different MP3 compression bit rates

VII. CONCLUSIONS

This paper proposes a technique to detect adulterations in digital audio recordings through the analysis of disturbances in ENF interfering signals with good results. The methods in [11] and [14] detect tampering from abnormal variations in ENF, caused by abrupt phase discontinuities due to suppressions or insertions of audio segments.

The proposed SPHINS method uses a ESPRIT-Hilbert ENF estimator that summarizes the ENF disturbances using the sample kurtosis as a measure of outlierness [21]. The computed kurtosis are vectorized and applied into a SVM classifier to indicate the presence of tampering.

To assess the proposed SPHINS method, the same database used in [11] and [14] is employed. Similar tests are performed evolving degradations due to noise addition and nonlinear saturation. The method is also evaluated when audio files are recoded with different MP3 compression bit rates.

The results show that the proposed method when applied to

TABLE VIII
SPHINS CROSS-VALIDATED ERROR RATES WITH CARIOCA 1 DATABASE RECODED WITH DIFFERENT MP3 COMPRESSION BIT RATES

BR(kbps)	OER(%)	FPR(%)	FNR(%)
16	5.95	6.4	5.5
32	5.45	5.2	5.7
64	5.1	3.1	7.1
128	4.5	2.7	6.3

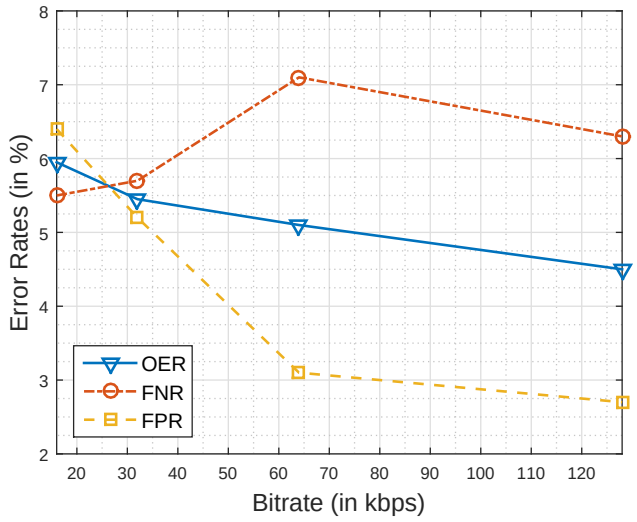

Fig. 14. SPHINS cross-validated error rates with Carioca 1 database recoded with different mp3 compression bit rates

the database in its original clean form performs a 4 % EER, with a 95.5 % of accuracy (4.5 % OER) in cross-validation tests.

The methods in [11] and [14] present performance degradation in the presence of noise and nonlinear saturation. Despite showing a similar behavior, the proposed SPHINS method gives better results than these techniques for low SNR regimes and for scenarios with nonlinear digital saturation, when applied to the same Carioca 1 database. Moreover, there is a small degradation in SPHINS performance to classify MP3 audio files compressed with low bit rates.

As future works, the application of the proposed approach to other corpora is considered. Moreover, the EER can be reduced further by combining SPHINS with methods that extract ENF from video image signal [13]. Another research line considered is to evaluate the effectiveness of the proposed SPHINS method in the tampering detection through other audio embedded narrow-band signals as, for example, the horizontal sync pulses in video signal.

Acknowledgments

The authors thank the Brazilian research and innovation agencies FAPDF (Research Support Foundation of the Federal District), FINEP (Agreement RENASIC / PROTO 01.12.0555.00), CAPES and CNPq under the FORTE Project - CAPES Forensic Sciences Announcement 25/2014 and the Program Science without Borders - Aerospace Technology supported by CNPq, CAPES for the postdoctoral scholarship abroad (PDE) number 207644/2015-2 and CAPES joint double degree scholarship number 88887.115692/2016-00

References

[1] R. Maher, "Audio Forensic Examination," *IEEE Signal Processing Magazine* , vol. 26, no. 2, pp. 84-94, Mar. 2009.

[2] X. Pan, X. Zhang, and S. Lyu, "Detecting Splicing in Digital Audios Using Local Noise Level Estimation," in *Proc. IEEE ICASSP*, pp. 1841-1844, 2012.

[3] H. Malik, "Acoustic Environment Identification and its Applications to Audio Forensics," *Trans. Inf. Forensics Security*, no. 11, pp. 1827-1837, 2013.

[4] R. Yang, Z. Qu, and J. Huang, "Exposing MP3 Audio Forgeries Using Frame Offsets," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 8, pp.35:1-35:20, 2012.

[5] Q. Liu, A. Sung, and M. Qiao, "Detection of Double MP3 Compression," *J. Cognitive Computing*, vol. 2, no. 4, pp. 291-296, 2010.

[6] H. Farid, "Detecting Digital Forgeries Using Bispectral Analysis," *AI Lab, Massachusetts Institute of Technology*, Tech Rep. AIM-1657, 1999.

[7] H. M. Sohaib Ikram, "Microphone Identification Using Higher-Order Statistics," in *Proc. AES 46th International Conference: Audio Forensics*, Jun. 2012.

[8] R. Tavora, F. A. Nascimento, "Detecting Replicas within Audio Evidence Using an Adaptive Audio Fingerprinting Scheme", *J. of the Audio Engineering Society*, vol. 63, no. 06, pp.451-452. Jun. 2015.

[9] S. Gupta, S. Cho, and C. C. J. Kuo, "Current Developments and Future Trends in Audio Authentication," *IEEE Multimedia*, vol. 19, no. 1, pp. 50-59, 2012.

[10] A. J. Cooper, "The Electric Network Frequency (ENF) as an aid to authenticating forensic digital audio recordings: An automated approach," *in Proc. AES 33rd Int. Conf. Audio Forensic, Theory and Practice*, Jun. 2008

[11] D. P. N. Rodriguez, J. A. Apolinário Jr., and L. W. P. Biscainho, "Audio Authenticity: Detecting ENF Discontinuity with High Precision Phase Analysis", *IEEE Trans. Inf. Forensics Security*, vol. 5, no. 3, pp. 534-543, 2010.

[12] A. Hajj-Ahmad, R. Gardi, M. Wu, "Instantaneous frequency estimation and localization for ENF signals", in *Proc. IEEE APSIPA ASC*, 2012.

[13] R. Gargi, A. L. Varna, M. Wu, "Seeing ENF: natural time stamp for digital video via optical sensing and signal processing", in *Proc. 19th ACM international conference on Multimedia*, pp.23-32, 2011.

[14] P. A. A. Esquef, J. A. Apolinário Jr., L. W. P. Biscainho, "Edit Detection in Speech Recordings via Instantaneous Electric Network Frequency Variations", *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2314-2326, Dec. 2014.

[15] R.O Schmidt, "Multiple Emitter Location and Signal Parameter Estimation," *IEEE Trans. Antennas Propagation*, vol. 34, no. 3, pp.276-280, 1986.

[16] R. Roy, T. Kailath, "ESPRIT - Estimation of Signal Parameters via Rotational Invariance Techniques", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 7, pp. 984-995, 1989.

[17] D.G. Manolakis, V. K. Ingle, and S. M. Kogon, *Statistical and Adaptive Signal Processing*, MaGraw-Hill, Inc., 2000.

[18] J. Short, D. G. Infield, L. L. Freris, "Stabilization of grid frequency through dynamic demand control", *IEEE Transactions on Power Systems*, vol. 22, no.3, pp:1284-1293, 2007.

[19] ANEEL, *Procedimentos de Distribuicão de Energia Eletrica no Sistema Eletrico Nacional-PRODIST*, mod. 8, 2013.

[20] P. D. Welch, "The use of Fast Fourier Transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms", *IEEE Transactions on Audio and Electroacoustics*, vol.15, no 2, pp. 70-73, 1967.

[21] D. Pea, F. J. Prieto, "Multivariate Outlier Detection and Robust Covariance Matrix Estimation", *Technometrics*, vol. 43, pp. 286-310, 2001.

[22] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[23] W. Liu, J. Príncipe, and S. Haykin, *Kernel Adaptive Filtering: A Comprehensive Introduction*. New York: Wiley, 2010.

[24] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance", in *Proc. Eur. Conf. Speech Communication and Technology*, Rhodes, Greece, Sep. 1997.

Paulo Max Gil Innocencio Reis received the Diploma degree in telecommunications engineering in 1998 from the Military Institute of Engineering (IME) in Rio de Janeiro, Brazil, and his M.B.A degree in Telecommunications Business in 2004 from the Federal Center of Technological Education (CEFET/RJ) in Rio de Janeiro, Brazil. Actually, he is a graduating student at University of Brasilia (UnB), Brasília, Brazil, since 2014. He is an Official Forensic Expert with National Institute of Criminalistics, Brazil, since 2006, working on audio, video and still image forensic analysis. He was the general-chair of the International Conference on Multimedia Forensics, Surveillance and Security (ICMedia'12) in 2012, and the local-arrangements chair of the IEEE International Workshop on Information Forensics and Security (WIFS'11) in 2011. His professional interests include digital signal processing, digital image processing, speech and audio processing, audio and image forensics.

João Paulo Carvalho Lustosa da Costa received the Diploma degree in electronic engineering in 2003 from the Military Institute of Engineering (IME) in Rio de Janeiro, Brazil, his M.Sc. degree in telecommunications in 2006 from University of Braslia (UnB) in Brazil, and his Doktor-Ingenieur (Ph.D.) degree with Magna cum Laude in electrical and information engineering in 2010 at Ilmenau University of Technology (TU Ilmenau) in Germany. Since 2010, he coordinates the Laboratory of Array Signal Processing (LASP) and since 2014, he works as a senior researcher at the Ministry of Planning in projects related to Business Intelligence. He was a visiting scholar at University Erlangen-Nuremberg, at Munich Technical University, at Seville University, at Harvard University, at Ilmenau University of Technology and at Fraunhofer Institute for Integrated Circuits IIS. Currently he coordinates a project related to distance learning courses at the National School of Public Administration and a special visiting researcher (PVE) project related to satellite communication and navigation together with the German Aerospace Center (DLR) supported by the Brazilian government.

He was a Guest Editor for Hindawi International Journal of Antennas and Propagation, special issue on "MIMO Antennas in Radar Applications 2016" and he served as the general-chair of the International Conference on Cyber Crime Investigation (ICCyber), International Conference on Multimedia Forensics, Surveillance and Security (ICMEDIA) and International Conference on Forensic Computer Science (ICoFCS) in 2015. He has been named an IEEE Senior Member in Signal Processing in 2015. He obtained five best paper awards on the following conferences: IV IEEE International Conference on Ultra Modern Telecommunications and Control Systems (ICUMT'12), ICoFCS'12, ICoFCS'13, ICoFCS'15, and 19th IEEE International Conference on OFDM and Frequency Domain Techniques (ICOF'16).

Ricardo Kehrle Miranda received the Diploma degree in telecommunications engineering in 2010 and the M.Sc. degree in electronic and automation systems in 2013 both from the University of Brasília (UnB), Brazil. In Germany, he did part of his M.Sc. work at the University of Erlangen-Nuremberg in 2012 and was as visitor student at the German Aerospace Center (DLR) in February and March of 2014. In Japan, he was an intern in the space industry for 8 months in a partnership with the Wakayma University, the Tokyo University and the Brazilian Aerospace Agency. Currently, he pursues a double Ph.D. degree at both the UnB and the TU Ilmenau with interests on array signal processing, multilinear algebra and space communications. His publications include 1 conference paper on microphone array signal processing, 3 conference papers and 1 journal paper on beamforming and rank reduction. .

Giovanni Del Galdo Giovanni Del Galdo studied telecommunications engineering at Politecnico di Milano. In 2007 he received his doctoral degree from Technische Universitt Ilmenau on the topic of MIMO channel modeling for mobile communications. He then joined Fraunhofer Institute for Integrated Circuits IIS working on audio watermarking and parametric representations of spatial sound. In 2012 he was appointed full professor at TU Ilmenau on the research area of wireless distribution systems and digital broadcasting. His current research interests include the analysis, modeling, and manipulation of multi-dimensional signals, over-the-air testing for terrestrial and satellite communication systems, and sparsity promoting reconstruction methods .